A Bioinformatics Pipeline for Recovering Misidentified Proteins

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Sudeep Mehrotra

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

To examine the response of wheat to different temperatures and photoperiods at the proteomic level, a series of experiments was performed at the University of Saskatchewan, College of Agriculture and Bioresources, Department of Plant Science. Tandem-mass spectrometry (MS/MS) was used for protein identification. The iTRAQ approach was used to generate raw data for protein quantification. The Pro Group protein identification software was used for protein identification and quantification of differentially expressed proteins. Despite the input samples being from a plant, the software reported non-plant proteins. The traditional approach used by scientists to deal with this problem is to use sequence alignment software to find close green-plant homologs of the non-plant proteins from a plant-only database. Such a technique is problematic since homology-based sequence similarity does not generally equate to similarity of mass spectra. In this work a more radical approach was investigated and implemented. A bioinformatics pipeline was designed and implemented to report plant proteins misidentified by the Pro Group software. The approach drew its idea from the fact that MS/MS-based protein identification uses peptide fragments/ions bearing unique m/z values in the mass spectra. From the reported non-plant proteins and associated peptides, putative m/z values of the peptides are generated and then used to find alternate hits from a green plant-only database. The pipeline uses three different heuristics, each generating a list of candidate proteins. The proteins reported consistently across the three reported lists have the highest likelihood to be present in the original sample. To evaluate the performance of the pipeline, three separate experiments were performed. A set of known plant peptides, a combination of known plant and non-plant peptides and a set of known non-plant peptides were used as input to the pipeline. For each experiment a stringency value (threshold value) was set by the user. Better results were observed by specifying a tighter stringency; that is, more plant proteins were reported consistently across the three reported lists. The research presented in this thesis shows that m/z values, consideration of unique peptides and accounting for proteins with shorter sequences can be used to identify proteins. These characteristics can be used to identify proteins when limited information is available, in this case a list of non-plant proteins reported as being present in a plant-derived sample. The information available was limited because the original input data was already processed by the Pro Group software. The approach presented here is an alternative to a wet lab scientist using sequence alignment tools, sequence databases, and homology-based search. The pipeline can be enhanced by adding various other modules. The results presented here could be used as a foundation for a further study.

Acknowledgements

I would like to thank my supervisor Dr. Tony Kusalik, Department of Computer Science, for his guidance and encouragement from the very early stage of this research as well as giving me extraordinary experiences throughout the research work. I would like to thank my co-supervisor Dr. Gordon Gray, Department of Plant Science, for giving me the opportunity to work on this project and for his guidance throughout the entire length of the research. I would also like to thank Dr. Andrew Ross, PBI-NRC, and Dr. Ian McQuilian, Department of Computer Science, for their advice, guidance, and suggestions. My family gave me consistent support through the years. This thesis would not be possible without their support.

This thesis is dedicated to my mother, whom I miss the most.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AA	Amino Acids
ASCII	American Standard Code for Information Interchange
R	Arginine
N	Asparagine
BLAST	Basic Local Alignment Tool
BLOSUM	Blocks of Amino Acid Substitution Matrix
BC	British Columbia
CPU	Center Processing Unit
C-terminus	Carboxyl-terminus
c	Charge
E-value	Expectation value
EMBOSS	European Molecular Biology Open Software Suite
ESI	Electrospray Ionization
ExPASy	Expert Protein Analysis System
FN	False Negative
FP	False Positive
G	Glycine
GB	GigaByte
GHz	Giga Hertz
HPLC	High Performance Liquid Chromatography
H	Histidine
IBM	International Business Machines
iTRAQ	Isobaric Tags for Relative and Absolute Quantitation
I	IsoLeucine
L	Leucine
Llimit	Lower Limit
K	Lysine
MS	Mass Spectrometry
m	Mass
MALDI	Matrix Assisted Laser Desorption Ionization
mRNA	Messenger Ribonucleic Acid
MW	Molecular Weight
MSDB	Mass Spectrometry Protein Sequence Database
MM	Mutation Matrices
NCBI-REFSEQ	National Center for Biotechnology Information-Reference Sequence
NMR	Nuclear Magnetic Resonance
N-terminus	Amino-terminus
PAM	Point Accepted Mutation
PBI-NRC	Plant Biotechnology Institute-National Research Council
PPM	Parts Per Million
PMF	Protein Mass Fingerprinting
P	Proline
PPV	Positive Predictive Value
PTMs	Post Translation Modifications
RAM	Random Access Memory
TAIR	The Arabidopsis Information Resource
TP	True Positive
TN	True Negative
ULimit	Upper Limit
Y	Tyrosine

Chapter 1

Introduction

The term proteome refers to the entire set of proteins produced by a cell, tissue or an organism. The study of proteomes is called proteomics [48]. One experimental technique to study proteins is mass spectrometry (MS), which is used to identify and quantify (investigate expression levels of) proteins at given locations and time. Protein levels change in response to changes in the internal and/or external environment. For example, changes in the physical environment, such as fluctuations in temperature and photoperiod are forms of external stress that affect protein levels in wheat plants [23]. Two other important aspects of proteomic studies are structural and protein interaction studies [48]. Structural studies involve use of techniques such as X-ray crystallography and/or nuclear magnetic resonance (NMR) to examine the final 3-D conformation of the proteins. Protein interaction studies examine a protein's interaction with other cellular complexes and how proteins interact among themselves (protein-protein interactions) [17, 48]. In this thesis, the focus is on functional studies and in particular, the use of MS for protein identification.

Mass spectrometry has emerged as the primary tool for high-throughput protein identification [32, 34]. Further, in combination with other available techniques, protein expression levels can be also examined [3]. Protein sequences are long chains of amino acids (AA) which can be broken down into short sequences called peptides. Various characteristics such as molecular weight (MW), charge state, mass-to-charge (m/z) ratio and relative abundance of peptides are used for protein identification and to determine protein expression levels, using techniques such as tandem-MS (MS/MS) [32, 42, 52, 53] and Isobaric Tags for Relative and Absolute Quantitation (iTRAQ) [3, 4]. Mass spectrometry is a high-throughput technique producing large amounts of data. Expert manual interpretation of the data is time-consuming and requires computational assistance to deal with large data sets containing thousands of spectra. Thus, protein identification software packages are typically used for protein sequencing/identification.

Protein identification and quantification are two essential steps in proteomic studies [43]. As the name suggests, protein identification refers to determination of proteins in a given protein sample. Protein quantification refers to determination of the level of (change in) protein concentration in response to changes in the internal and/or external environment, at a given location and time [47]. Protein expression level studies can be performed without protein identification. However, in order

to understand the complete biological functions of proteins, identifying the proteins of interest becomes essential.

One aspect of plant proteomics involves the study of changes in the expression levels of plant proteins. Plants respond differently to varying stress. Stress caused by changes in temperature and amount of light is of particular interest to the researchers at the University of Saskatchewan in the Department of Plant Science. Four different genotypes of wheat were selected for their study. Thirteen different experiments were performed. In each experiment the plants were grown under high and low temperatures as well as short and long photoperiods. Following growth and harvesting, chloroplasts were isolated and purified. The iTRAQ approach [18] was used to generate raw data for protein expression analysis and tandem-MS technique was used for protein identification. The Pro Group software from Applied Biosystems[1] was used for protein identification and quantification in the samples. For each experiment a Pro Group report was generated. An example of a Pro Group report (information such as accession number of identied proteins, protein names, two scores used by the software) is shown in Figure A.1 in Appendix A. Unexpectedly, from the known plant samples, a number of non-plant proteins were reported by the software along with plant proteins. This observation motivated the research described in this thesis.

This thesis contains five chapters in addition to this one: Background, Methodology, Results, Discussion and Conclusions and Future Studies. The Background chapter provides details on the workings of MS for protein identification. The chapter covers details on how the Pro Group software was used for protein identification. Furthermore, possible reasons for the reporting of non-plant proteins by the (Pro Group) identification software are also discussed.

The Methodology chapter begins with the discussion of various assumptions which were necessary to complete the research. Various important concepts utilized in the design of the pipeline are also discussed. Next, details are presented on the methodology that is used for the recovery of plant proteins from the set of non-plant peptides. Within the Methodology the Testing section discusses various test cases which were used to examine aspects of the pipeline and analyze the results obtained. The Testing section provides details on how the pipeline was tested, and provides a brief description on how results were interpreted. The output from various test cases are presented and the performance of the pipeline is discussed.

The Results chapter presents the output from the pipeline in great detail. Three different experiments were conducted, whose results are shown.

The research presented in this thesis shows that with minimal information at hand, which is knowledge of peptide sequences, a reasonable attempt can be made to recover the misidentified proteins. Initially, to recover plant proteins a methodology using physciochemical based scoring matrix along with a plant database was proposed. Details of this (abandoned) methodology are

[1] http://www.appliedbiosystems.com

presented in the Discussion chapter. Furthermore, during the design of the pipeline alternative ideas were explored but not adopted in the final version of the pipeline. Details of these ideas are also presented in the Discussion.

The Conclusions and Future Studies chapter discusses the shortcomings of the designed pipeline and suggests ideas that could be implemented to overcome these limitations.

1.1 Thesis Objective

Report of non-plant proteins from a sample of plant proteins signifies a limitation of protein identification by the Pro Group software. When this occurs, many researchers use sequence databases and sequence alignment tools to find close plant homologs of the reported non-plant proteins. There are limitations to this approach.

In this thesis, an alternative bioinformatics approach is presented. The novel pipeline attempts to use the non-plant peptides reported by the original software for protein identification. The pipeline was implemented with the main objective to report plant proteins potentially misidentified by the original software. Mass-to-charge ratios of all the non-plant peptides were used by the pipeline to identify the plant proteins using a plant-only database. The pipeline used three different heuristics, each generating a list of candidate proteins. The proteins reported consistently across the three lists had the highest likelihood to be present in the original sample.

The pipeline was tested and the results were evaluated. The testing of the pipeline involved investigation of different input parameters and collection of data for error statistics to evaluate the response of the pipeline in response to changes in the parameters. The pipeline was initially tested with know input data and then experiments were conducted with real data sets (data reported by the software).

Chapter 2

Background

Use of MS along with database search has become a widely accepted method for protein identification and quantification. However, even with the emergence of improved algorithms and carefully curated databases, problems remain with the quality of protein "hits" returned by the algorithms [10, 11]. Research conducted elsewhere [6, 10, 11, 30] has reported various reasons for these problems. For instance, choice of, and consistency in, parameters for searching the database, the kind of database used, incompleteness of information in the database, the way data-sets are obtained (use of different mass spectrometers and methods), and bias of algorithms towards a certain class of protein(s) or protein(s) with higher mass are a few examples of factors that can hamper the correct identification of proteins [11]. Accuracy of identification software can be improved by using additional parameters, such as specifying increased numbers of post-translation modifications and missed cleavage sites, but at a cost of negatively affecting performance. The specifics of these parameters are described in the next section. If the source of the protein sample-set is known, it is expected that the identification software would identify proteins from the same organismal source. Report of proteins from other than the known source requires further investigation. For example, if proteins are extracted from plants, it is expected that the proteins identified would be from plant species. Any deviation would encourage investigation and resolution of the problem.

In our plant proteomics research, Pro Group software was used to identify proteins in a sample and their expression levels. When the data generated by Pro Group was analyzed, it showed many non-plant proteins (Figure 2.1). Report of non-plant proteins from samples of plant proteins indicated that further investigation was necessary to ascertain why such proteins were reported and, more importantly, to recover any unidentified plant proteins. A new method is proposed in this work to provide a list of potential plant proteins that might have been reported by Pro Group.

2.1 Mass Spectrometry

Mass spectrometry has become a vital tool for proteomic studies [2]. The basic architecture of a mass spectrometer consists of three parts: a source of ions, a mass analyser and a detector. Electrospray ionization (ESI) or matrix-assisted laser desorption/ionzation (MALDI) are the two

AF237621 NID: - Homo sapiens		1	EQIQSLNNQFASFIDKVR
Keratin, type II cytoskeletal 1 (Cytokeratin 1) (K1) (CK 1) (67 kDa cytokeratin) (Hair alpha protein		99	GGGGGGYGSGGSSYGSGGGSYGSGGGGGGGR
AF304164 NID: - Homo sapiens		86	SLNNQFASFIDJ
keratin 1, type II, cytoskeletal - human			

Figure 2.1: Representative example to show a non-plant protein reported by the Pro Group software. In this example, Pro Group reports a "hit" from a human (*Homo sapiens*) protein. The highlighted text (blue background) on the right side lists the peptides associated with the protein name "AF23721 NID: Homo Sapiens". The figure only shows a portion of the full output.

most commonly used techniques for generating ions [2, 3]. Both volatilize and ionize (provide a charge to) peptides for MS analysis. One major difference between the two techniques is that ESI produces multiple ions of peptides, each ion with a different charge, while single charge ions are predominant in MALDI. A mass analyzer measures the m/z ratio of the ionized peptides. The counts of these charged ions are then recorded by the detector.

In MS, protein identification is a multi-step process that involves use of mass spectra and database search algorithms (Figure 2.2). A peptide sequence is a short chain of AAs.

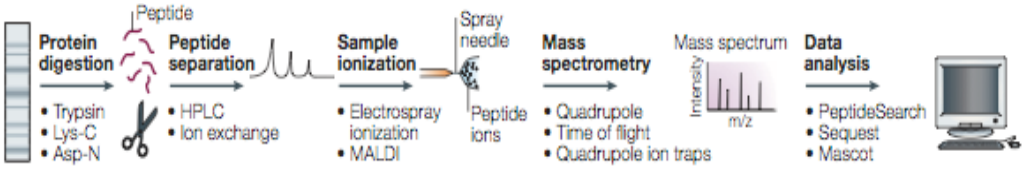

Figure 2.2: Steps in protein identification using MS. Protein samples are collected and digested using site-specific proteolytic enzymes to produce peptides. The peptides are separated using various techniques and ionized by using electrospray ionization or matrix-assisted laser desorption/ionization. Different MS techniques can then be used to record m/z ratios. The recorded m/z ratios are used by different search engines for protein identification. Figure modified from [42].

A protein is a polymer of peptides. In order to break long chains (polypeptides), different proteolytic enzymes (proteases) are used. These enzymes are site-specific; i.e. the location where enzyme will cut is known. Trypsin is the most commonly used protease. It cleaves a protein on the carboxy-terminal side of arginine (R) and lysine (K) not followed by a proline (P) residue (Figure 2.3). Using chromatography techniques followed by ionization, peptides are and introduced to the mass spectrometer to generate mass spectrum.

A mass spectrum is a distribution of ions. The two axes of a mass spectrum are m/z (x-axis) and intensity (y-axis) (Figure 2.4). The intensity (counts) and m/z ratios (peaks) of the peptide ions are recorded. This information is used for protein identification.

Protein identification software performs an *in silico* digestion of protein sequences that are stored in a database. The software then matches the observed spectra to the calculated spectra

(1)
>YP_874732.1 synthase CF0 subunit I [Agrostis stolonifera]
MENVTHSFVFLAHWPSAGSFGLNTDILATNLINLTVVVGVLIFFGKGVLKDLLDN
RKQRILSTIRNSEELRRGTIEQLEKARIRLQKVELEADEYRMNGYSEIEREKANLIN
ATSISLEQLEKSKNETLYFEKQRAMNQVRQRVFQQAVQGALGTLNSCLNTELHF
RTIRANIGILGSMEWKRKLN

(2)

Cterm	Nterm	Sequence
.	G	MENVTHSFVFLAHWPSAGSFGLNTDILATNLINLTVVVGVLIFFGK
R	T	VFQQAVQGALGTLNSCLNTELHFR
R	.	ANIGILGSMEWKRKLN
K	S	ANLINATSISLEQLEK
K	M	VELEADEYR
R	E	MNGYSEIER
K	Q	NETLYFEK
K	N	QRILSTIR
R	A	GTIEQLEK
R	G	NSEELRR
K	V	ARIRLQK
K	Q	DLLDNRK
R	Q	AMNQVR
K	D	GVLK
R	A	TIR
R	V	QR
K	A	QR
R	A	EK
K	N	SK

Figure 2.3: Predicted peptide digest of a protein sequence. The given sequence (1) is of a plant protein that is 186 AA long. If the protein was digested using trypsin, 19 peptides/fragments would be produced (2). The C-terminus of the peptides is either R or K. The symbol "." represents a null value either before the start or after the end of the protein sequence. Cterm and Nterm represent the two terminals of the peptides and the values under these two headings are the AAs present at the two terminals. For example, consider the last peptide sequence SK. In the actual sequence the fragment would be KSKN. The data was collected by using the application digest from the EMBOSS set of programs, version 4.0.0 [35]. Part (2) of the figure shows the output (verbatim) given by digest.

of peptides from the database and assigns a score to the matched spectra. The score reflects the confidence level in matching mass spectra of the observed sequence (query sequence) with the peptide sequences (calculated sequences) present in the database. Proteins are identified when the query peptides are matched uniquely with the peptides from the database. The reported/identified proteins should have higher scores. This method is called peptide mass fingerprinting (PMF). It is important to note that search engines match MS spectra with peptide masses from the database and not to peptide sequences. Achieving good quality matches and correct identification of proteins remains a challenge [6, 10, 11, 30].

Various MS instruments are available for proteomic studies. Variations are made by making additions to the basic architecture for higher quality analysis. Such modifications are, for example, necessary to identify proteins that are present at only trace levels in complex samples. Another modification is the addition of another mass analyzer in tandem, called tandem mass spectrometry (MS/MS). After recording the m/z ratio in the first mass analysis, selected peptide (ions selected in real time or within a pre-selected range) ions are isolated and passed to the second mass analyzer.

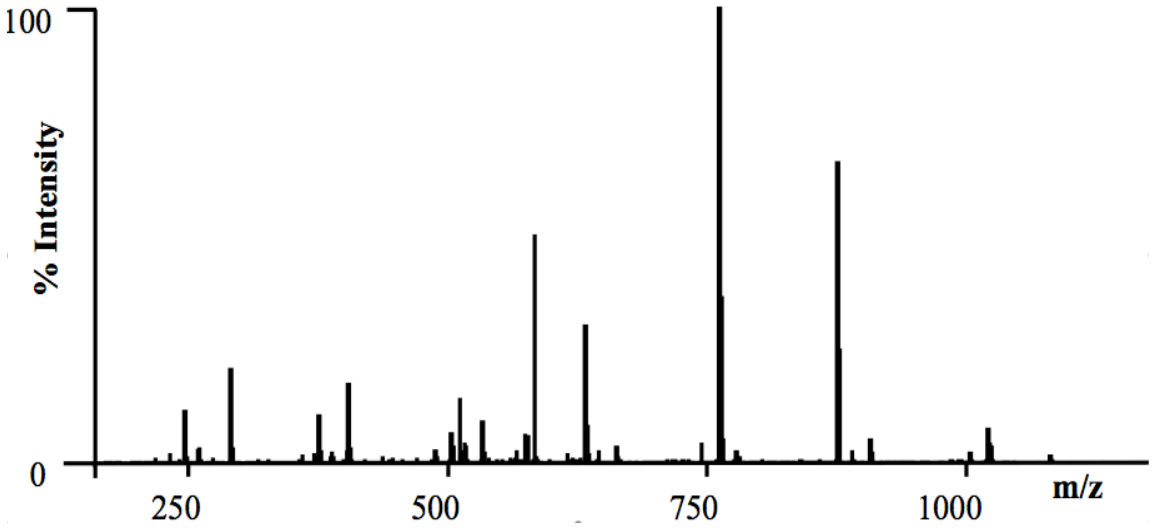

Figure 2.4: A typical mass spectrum for an unknown protein digest. On the x-axis are the m/z ratios and the y-axis shows their relative intensity. Every peak represents a peptide ion. The peaks are interpreted by various protein identification software. Figure modified from [20].

Peptide fragment ions are generated and recorded (Figure 2.5). The ions at the first mass analyzer are called parent or precursor ions and ions at the second stage are called daughter ions (Figure 2.6). From the mass of the parent ions and mass spectra of the daughter ions, the peptide is infered. Given identification of multiple peptides from a given protein, the protein can then be determined.

2.1.1 Protein Identification using MS/MS

The premise of peptide mass fingerprinting (PMF) is that each peptide and its fragments have a unique signature/fingerprint, i.e. set of m/z ratios. Further, if the m/z ratios and charge of the parent ions and m/z ratios of the daughter ions are known, search algorithms can then be used for identifying the peptides, and hence, proteins. The basic idea behind a search algorithm is to correlate the observed peaks (coming from peptides and their fragments) with the set of peaks generated after *in silico* digestion of the protein sequences in the database (Figure 2.7).

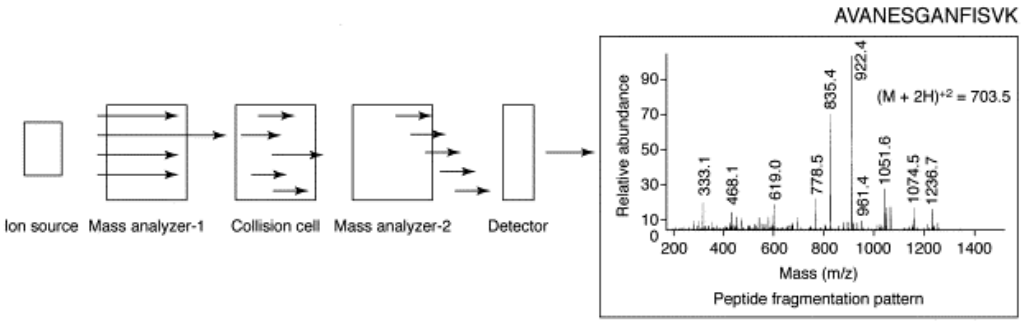

Figure 2.5: Tandem-MS architecture. The process of MS/MS involves the use of two mass analyzers. At the ion source, peptide ions are generated and passed to the first analyzer, mass analyzer-1. Selected ions (parent ions) of a specific m/z are then passed through a collision cell. The resulting fragment ions (daughter ions) are counted and recorded by a second mass analyzer, mass analyzer-2, to produce a tandem mass spectrum. The pattern/mass spectrum shown in the figure is for the peptide with AA sequence AVANSEGANFISVK produced at mass analyzer-2 for a parent peptide selected at mass analyzer-1 with MW 1407 Da (approximately). The m/z value at mass analyzer-1 is 704.5 by addition of two protons (shown as $(M+2H)^{+2}$ in the figure). Figure modified from [53].

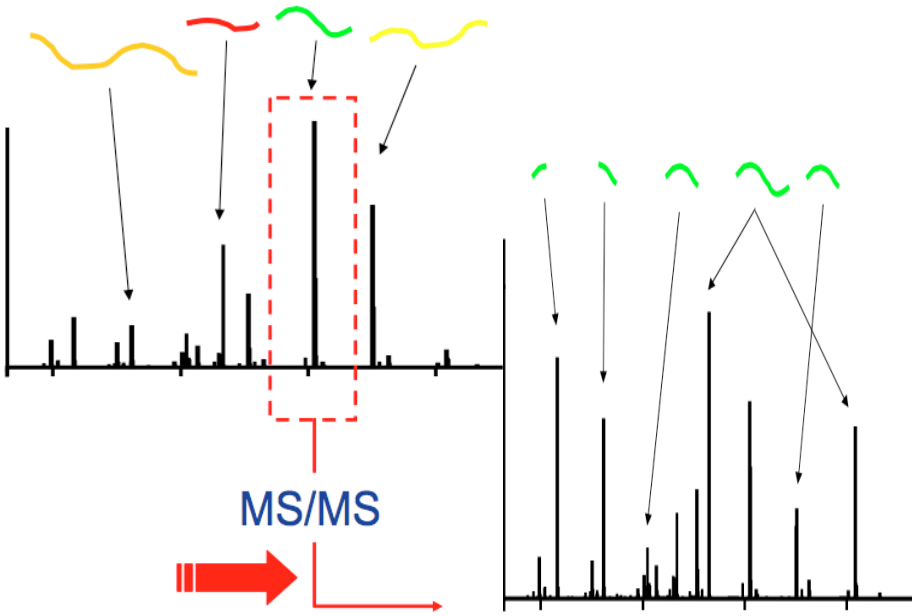

Figure 2.6: Mass spectra displaying parent and daughter ions. In MS/MS, a precursor peptide (parent) (shown in green color) is selected at the first mass analyzer. Daughter ions in the second mass analyzer are determined and recorded. Figure modified from [20].

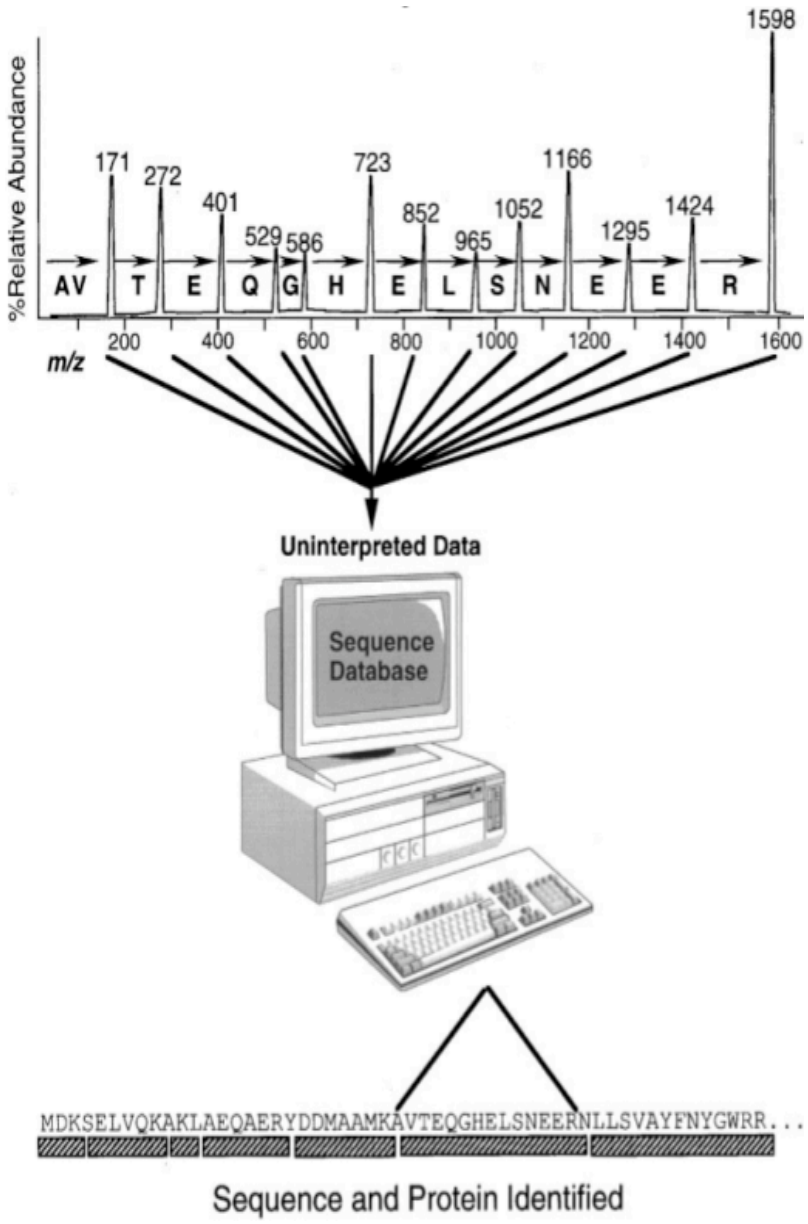

Figure 2.7: Use of uninterpreted peptide MS/MS data and database search software for protein identification. The x-axis shows the m/z ratios of peptide fragments obtained using MS/MS. The y-axis shows the relative intensity of the ions. The mass spectrum shows peaks, where each peak corresponds to a daughter ion. For example, the mass spectrum shows that ion AV has m/z of 171. Search algorithms correlate this information with *in silico* digested peptides in the database, find a reasonable match (based on a score), and report the identified protein. Figure modified from [52].

There has been a tremendous increase in the availability of search engines for peptide and protein identification from mass spectra. Protein identification search engines are all distinct, and they differ in various aspects such as the scoring schemes, the set of parameters provided to perform the search, and options to select different databases. Beavis and Fenyo [6], Nesvizhskii and Abersold [30] and Vlhinen [50] have discussed and compared features of the different search algorithms most commonly used for correlating the MS/MS data with a protein sequence database. Availability of different protein identication software programs and sequence databases have greatly influenced the proteomics field. However, there are various limitations associated with these methods affecting the identication of the proteins in the sample.

2.1.2 iTRAQ

In proteomic studies, apart from protein identification, scientists are also interested in measuring the expression levels of proteins. iTRAQ is one of the techniques used with MS for protein quantification [12]. The basic technique involves site-specific labelling of digested peptides in such a way that the tagged peptides can be distinguished in the MS/MS mass spectrum. The tagged peptides produce signature peaks in the MS/MS output and the intensity of these peaks correspond to the counts/abundance of the peptides in the sample [39]. The ratios of peak areas (of the signature peaks) reflects the relative abundances of the peptides and the proteins in the sample. By using a reference standard sample, absolute quantification can be achieved [40]. For more information about iTRAQ refer to the papers by Chong et al. [12] and Ross et al. [37].

2.2 Protein Identification using Pro Group

There are many computational tools available for protein identification. For more information about various tools refer to Matthiesen [28]. For the experiments conducted at University of Saskatchewan in the Department of Plant Science, identification and quantification of the proteins in the samples were done by a combination of applications that come with the Pro Group software. Protein identification was performed by Pro ID and quantification by Pro QUANT. Results from both were then analyzed, filtered, grouped and displayed as a report by Pro Group. Selective information on Pro ID and Pro Group is provided in the next two sections. More information is described in the manual accompanying the Pro Group software [8]. The bioinformatics pipeline described in this thesis makes use of data from Pro ID and Pro Group. It does not use information from Pro QUANT.

2.2.1 Pro ID

Pro Group uses Pro ID for protein identification. The protocol for selecting the parent and daughter ions was set by scientists at the University of Victoria Genome BC Proteomics Center[1]. The list of observed spectra is matched with the theoretical (predicted) spectra from *in silico* digestion with trypsin of the MSDB database (Mass Spectrometry Protein Sequence Database, 2005)[2], not restricted to a particular taxonomy (i.e. all species). The quality of peptides identified is based on two parameters, *score* and *confidence*, which are described in detail in the manual accompanying the software [8]. Briefly, for each match between the two spectra (observed and theoretical), a *score* is calculated and assigned. The *score* is based on the number of matches of most intense peaks in the observed data with the theoretical. To calculate the *confidence*, the algorithm further incorporates two values, a *distance score* calculated for the peptides and a *total count of matches* returned after the search. A *distance score* can be described as a scale that ranks bins (collection of sorted peptides in various groups based on m/z values) according to *score*. Matches with high *scores* are reported as potential candidate peptides. The detailed description of various schemes used by the software is beyond the scope of this thesis. For more detail refer to the Pro Group manual [8].

There are various factors that can influence a score; for example, the threshold values selected for the tolerance, size of the theoretical peptides and post-translation modifications. Threshold values for MS and MS/MS refer to the maximum difference that can be allowed between an observed peptide and a theoretical peak. Specifying a large window size (tolerance) will increase the chances of random matches, thus reducing discrimination between the true peptides identified in a sample and false positives. However, lowering the threshold value too much also has a negative effect because valid matches can be missed [33]. The length of the peptide also plays a vital role. A longer peptide will have more ion fragments, so a match to this peptide will have a better score than a correspondingly good match to a shorter peptide. Also, AA sequences can undergo various modifications. The mass of the fragment or fragment ions are affected by such modifications.

The following example illustrates the workings of Pro ID. Assume that a precursor with a m/z ratio of 730 is selected, and the peptide sequence representing this m/z is broken into daughter ions (with corresponding observed peaks). If the peptide is composed of the AA sequence ABCDE, it can be broken down into various combinations of its constituent characters (AAs), such as: A, A + B (A and B), A + B/C/D (A and B or C or D), AB + C/D (AB and C or D), ABC + E (ABC and E), B, B + C (B and C), BC + D (BC and D), etc.
All the protein sequences in the database are *in silico* digested by the software and the peptides are

[1] http://www.proteincentre.com/home
[2] MSDB database as compiled by the Proteomics Group at Imperial College London, http://csc-fserve.hh.med.ic.ac.uk/msdb.html.

sorted into bins based on m/z values. Assume all the observed peptides with m/z of 730 will be in bin1. All the peptides in this bin are now further fragmented (corresponding to the second stage of MS/MS) and these fragments are than used for identification. The daughter ions in the respective bins are the candidate ions used for protein identification. A *score* is assigned based on the match between observed and the theoretical peaks (observed after *in silico* digestion of the database). Pro ID captures all the peptides based on matches between observed and theoretical spectra. Each match is given a *score*. Pro Group then finds justification (data) from other algorithms within the Pro Group software for such peptides and uses them for protein identification.

2.2.2 Pro Group

Pro Group acts as a second layer in protein identification. Results from Pro ID are the inputs for Pro Group. The scores given to peptides by Pro ID are not utilized at this stage. Pro Group's functionality includes:

- Determination of confidence scores for all proteins,
- Grouping of similar and redundant proteins,
- Generation of reports in various ready-to-use and export formats.

Pro Group determines a level of confidence by identifying proteins in context of other candidate proteins based on the results from Pro ID. The premise followed by Pro Group is: once a peptide is used for identification of a protein, the same peptide cannot be used again for other protein identification. Two *ProtScores* are calculated by Pro Group to enforce this condition for every protein identification. The two *ProtScores* are *Total ProtScore* and *Unused ProtScore*. *Total ProtScore* for a given protein is calculated by considering all the peptides pointing to that protein. It is an intermediate value that does not reflect confidence for the identification of the proteins. *Unused ProtScore* refers to only those peptides that are unique; i.e., spectra of peptides which are not already claimed by other putatively identified proteins. This score is the key to protein identification. Peptides contributing to a high *Unused ProtScore* are used for protein identification [8]. The difference between the two scores depends on the number of peptides which are allowed to contribute to the score. Figure 2.8 illustrates with an example the manner in which proteins are identified and justified by Pro Group. The idea behind Pro Group's strategy is to minimize false positives.

Occurrences of multiple entries (redundant sequences) in the reference database for the same protein sequence and close homologs can increase the rate of false positives [6, 10, 11]. Pro Group tries to distinguish such proteins and place redundant proteins in groups. Grouping of proteins by Pro Group is based on shared MS/MS peaks.

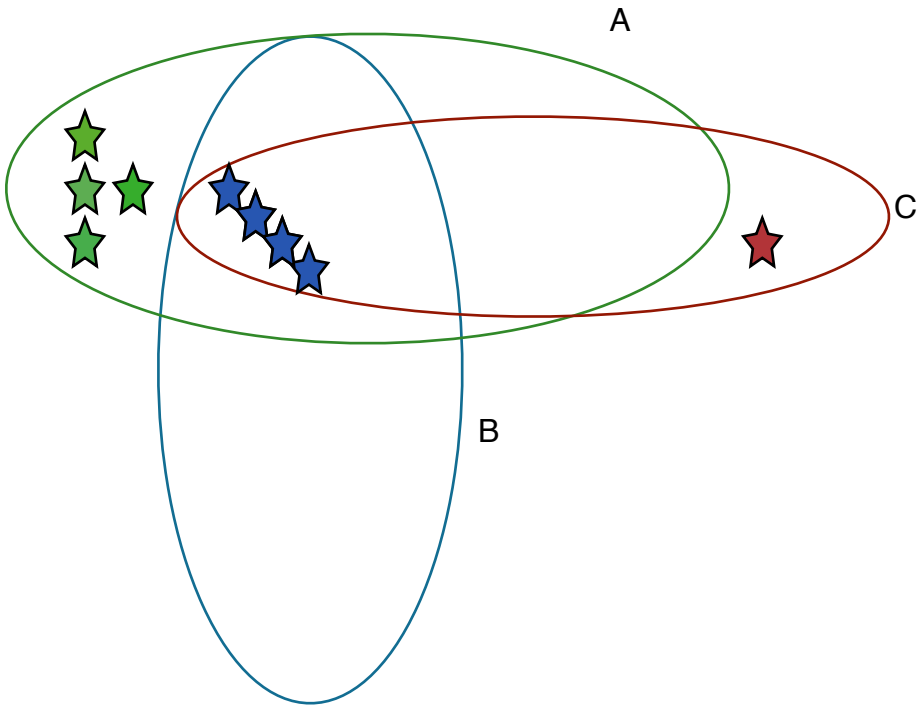

Figure 2.8: Grouping of proteins by Pro Group. In the figure ovals labelled A, B and C are the proteins returned by Pro ID, sharing colored stars representing peptides (evidence). Eight peptides are claimed for protein A (green+blue), 4 for protein B (blue) and 5 for protein C (blue + red). Since 8 peptides are already associated with protein A, the *Unused ProtScore* will be higher for this protein when compared to that of proteins B and C based on the scoring scheme used by Pro Group. Protein A will be declared as the "winner of the group" [9]. Figure modified from [9].

For the user's convenience, the justification for protein identification based on peptides, relevant scores and various other thresholds and parameters such as accession number, protein name and species name are presented in a report format. Further, a data export facility is provided to better manage, store and retrieve the data.

2.3 Sources of Errors and Potential Reasons for Non-plant Hits from Pro Group

Protein identification using MS-based approaches is a complex process. Much processing is required before proteins can be declared as identified in a sample, and there are various sources of error that can make true identification of proteins a challenging task. Errors can occur anytime between when a sample is extracted to when peptides from a reference database are used for protein identification. Various measures are taken during the whole process to help achieve higher quality results. It is assumed that the protein samples used in the original experiment are free from the most commonly observed contaminant, keratin. This section discusses specific sources of error that

can affect analysis by Pro Group for protein identification.

In tandem-MS, m/z ratios of the parent ions and m/z values of subsequent daughter ions are considered for protein identification using sequence databases. Errors can occur for various reasons during identification. The following example demonstrates how difference in total count of parent ions, incorrect correlation of daughter m/z values, and incorrect m/z values of parent ions can affect protein identification. Suppose a known plant protein is cleaved using trypsin and a set of 10 predicted peptide m/z ratios are produced (assuming z is +1 in m/z):
200.0848, 201.3708, 202.1186, 201.3896, 206.8737, 208.6343, 219.2555, 223.6406, 254.6446, 257.8059.
Further, suppose that using MS, the following peptide m/z ratios are actually produced:
200.0848, 201.3708, 207.5467, 208.6343, 219.2555, 223.4534, 240.4500, 256.3456. The experimentally determined m/z ratios are not complete (difference in the total count of peptides). Ten peptide m/z values were expected; however, 8 are observed. There are two missing m/z values. These missing values together or in combination of other m/z values could have been used by Pro Group for plant protein identification. The two missing values could be a result of various errors such as instrumental error, miss-cleavage sites or an incomplete mass spectrum. Furthermore, 4 observed m/z values (207.5467, 223.4534, 240.4500, 256.3456) are not related to the predicted m/z values (incorrect m/z values). Use of these m/z ratios by the software could return a non-plant hit. Ideally, the 4 matching m/z values (200.0848, 201.3708, 208.6343, 219.2555) can be correlated to a plant protein from the sequence database by Pro Group. However, due to sources of error discussed below it is possible that of the 4, three m/z values correlate to a non-plant protein (e.g. non-plant peptides could have identical m/z values) and only 1 to a plant protein. The software will give preference to the highest count, i.e. 3, resulting in report of a non-plant protein in the final output. Hence, missing m/z values, incorrect correlation, or incorrect m/z values can affect correct identification of plant proteins.

Discussion of all possible sources of error is not within the scope of this thesis. Only the most common errors such as contamination, PTMs, miss-cleavage sites and inclusion of isotopes are briefly discussed. In principle, in a typical mass spectrum each putative AA combination should have a peak and it should be easy to disambiguate them (from background noise, presence of isotopes and from each other). Mass spectra of a foreign particle (contaminant) can be quite similar to a true protein present in the sample. This can affect the identification, and non-plant hits can be reported. Other possible contaminants include viruses and bacteria.

Post-translation modifications are covalent events that can change the properties of a protein by addition of a modifying group to one or more AAs [27, 31]. Phosphorylation and methylation are examples of such modifications. PTMs can interfere with correct identification of protein sequences. Fixed modifications and terminal modifications are the two known types of PTMs that affect protein identification. A fixed modification is one that is found on all instances of a given

residue in a protein, while terminal modifications are found exclusively at the termini of a protein [54]. In the above example, an expected *m/z* peak of peptide was 223.6406. Suppose, however, that from the instrument 223.4354 was observed. The shift in the ratio could be attributed to a PTM resulting in incorrect protein identification.

Long stretches of proteins are digested into shorter, more comprehensible sequences by using site-specific proteolytic enzymes. Miss-cleavages occur when an proteolytic enzyme (trypsin in our case) skips cleavage site(s), resulting in different digestion of proteins and an unexpected set of peptide *m/z* values. Different digestion of proteins can also occur within the *in silico* digestion. This is illustrated with an example. Two software programs were used to digest a protein sequence selecting trypsin as the proteolytic enzyme. The programs reported different results. Eighteen peptide fragments are predicted by the digest program (Figure 2.3). However, digestion of the same protein sequence using the program PeptideCutter[3] produces a different set, containing 26 peptide fragments (Figure 2.9). Different sets of rules were followed by the programs for protein digestion, resulting in different peptide fragments. Differential digestion of the protein sequence will produce a variant theoretical mass spectrum.

Amino acids are composed of the elements hydrogen, carbon, nitrogen, oxygen, and sulphur. During MS/MS, a mass spectrum is also populated with peaks corresponding to naturally occurring isotopes of these elements. Variations in *m/z* ratio of peptides can be attributed to the presence of such isotopes. These variations can then interfere with the correct identification of the ions/peptides. Deisotoping is the process through which the most abundant (standard) isotopes are identified by the software. Monoisotopic masses of peptide ions are than determined.

To identify proteins from MS/MS data, sequence databases are used by search engines looking for correlations between observed spectra and spectra from *in silico* digestion of peptides in the database. Correct correlation, hence identification, can be hampered by problems in databases. Sequencing errors can result in presence of erroneous protein sequences in the database. Such sequences can compromise the results from search algorithms by leading to production of erroneous matches. Often databases are incomplete; that is, the true matching protein sequence does not exist in the database. Unfortunately, the search algorithms return a match for the input spectra regardless of whether the result is artificial, i.e. biologically not relevant. Hence, if the database is incomplete, a false positive may be produced. New high-throughout technologies allow new sequences to be identified ever more rapidly, reducing the incidence of false positives as a result of database completeness.

Various search engines for protein identification are available. Every search engine is different, considering – among other aspects – their scoring schemes. A *score* represents a confidence level in the prediction/protein identification by the software. In the case of Pro Group, for scoring and

[3] PeptideCutter is available at http://www.expasy.org/tools/peptidecutter/

(1)
>YP_874732.1 synthase CF0 subunit I [Agrostis stolonifera]
MENVTHSFVFLAHWPSAGSFGLNTDILATNLINLTVVVGVLIFFGKGVLKDLLDN
RKQRILSTIRNSEELRRGTIEQLEKARIRLQKVELEADEYRMNGYSEIEREKANLIN
ATSISLEQLEKSKNETLYFEKQRAMNQVRQRVFQQAVQGALGTLNSCLNTELHF
RTIRANIGILGSMEWKRKLN

(2)
MENVTHSFVFLAHWPSAGSFGLNTDILATNLINLTVVVGVLIFFGK
GVLK
DLLDNR
K
QR
ISTIR
NSEELR
R
GTIEQLEK
AR
IR
LQK
VELEADEYR
MNGYSEIER
EK
ANLINATSISLEQLEK
SK
NETLYFEK
QR
AMNQVR
QR
VFQQAVQGALGTLNSCLNTELHFR
TIR
ANIGILGSMEWK
R
K

Figure 2.9: Peptide maps generated by PeptideCutter showing all possible peptides (2) of a protein (1). The *in silico* digest produced is for the same protein as used in Figure 2.3. When a protein of interest is cleaved with an enzyme, miss-cleavage sites can effect experimental peptides lists by producing a different set of peptides. An alternate *in silico* tryptic digest produced 18 peptides (see Figure 2.3). However, in this case 26 peptides are produced for the same protein. This demonstrates that different set of rules were used for the digestion of the protein for the same proteolytic enzyme (trypsin).

identification purposes, the peaks (parent and daughter ions) are grouped or "binned" using m/z ratios from the current scan. These (observed) ions in each bin and their clusters are then correlated with theoretical (predicted) ions from a sequence database to identify proteins. For example, consider a hypothetical bin with the following observed data: m/z ratio of a parent ion is 393.1245; m/z of daughter ions are 280.1526, 213.1021,195.1523 and 182.0582. In an ideal case, Pro Group will correlate ions with exact matches from the database and report a peptide sequence. However, problems may occur resulting incorrect identification. For example, in this case, a combination of observed m/z values might correlate to non-plant peptide ions and, depending on threshold settings, a non-plant peptide can be declared as identified. Unfortunately, errors can also occur during binning of ions due to instrumental errors. An ion can be wrongly binned and this would lead to incorrect matches. Sharing of multiple peaks by several peptides from close homologs of proteins present in the sample or database further complicates the matching/identification process.

Fundamental biochemical properties are also the source of potential errors. For instance, the mass of a combination of ions can be nearly equivalent to mass of a single AA. For example, the mass of two glycines (G) is very similar to that of asparagine (N) (G 57.02146 + G 57.02146 $\simeq$ N 114.04293). The mass of isolucine (I), 113.08406, and that of leucine (L), 113.08406, are identical. Thus, a peak in the mass spectrum can be ambiguous, hampering the exact sequencing of peptide, thus affecting identification of proteins. Furthermore, correct identification of proteins can also be hampered by interfering peak(s) (peak(s) not belonging to the fragmented peptides) interfering with the correct analysis. Interfering peaks can occur due to contamination, instrumental error, stray ions or background noise [45].

Chapter 3

Methodology

Tandem-MS along with sequence databases have become a mainstay for proteomic studies. This methodology was used to identify proteins from samples collected from wheat. Non-plant proteins were reported by the Pro Group software. A novel methodology was designed to use the m/z ratios of non-plant peptides to recover potential plant proteins using a plant protein sequence database. The methodology described in this thesis is based on some key facts, concepts and assumptions.

3.1 Concepts and Assumptions

The premise for protein identification using MS/MS is the following: each protein has a set of signature peptide fragments cleaved by site-specific enzymes, and bearing unique m/z ratios. These signature m/z ratios of multiple peptides are used to identify proteins from databases. The m/z ratios of the plant and non-plant peptides is the main attribute that is used in the bioinformatics pipeline presented in this thesis. The pipeline accepts a set of non-plant peptides (collected manually from a Pro Group report). The pipeline then calculates the m/z ratios of each peptide and correlates them with the m/z ratios of *in silico* digested peptides from a plant-only database and reports a set of potential plant proteins.

Every protein reported by Pro Group is based on identifiable peaks in the mass spectra. All the plant protein peaks are assumed to be true/valid and correctly identified.

3.2 Implemented Methodology

The implemented methodology is based on the assumption that m/z ratios of peptides are unique and that the charge state(s) of the peptides in the original mass spectra can be predicted. The m/z ratios of non-plant peptides are correlated with m/z values of plant peptides obtained from a plant protein database to produce a list of plant proteins potentially misidentified by the Pro Group software (Figure 3.1).

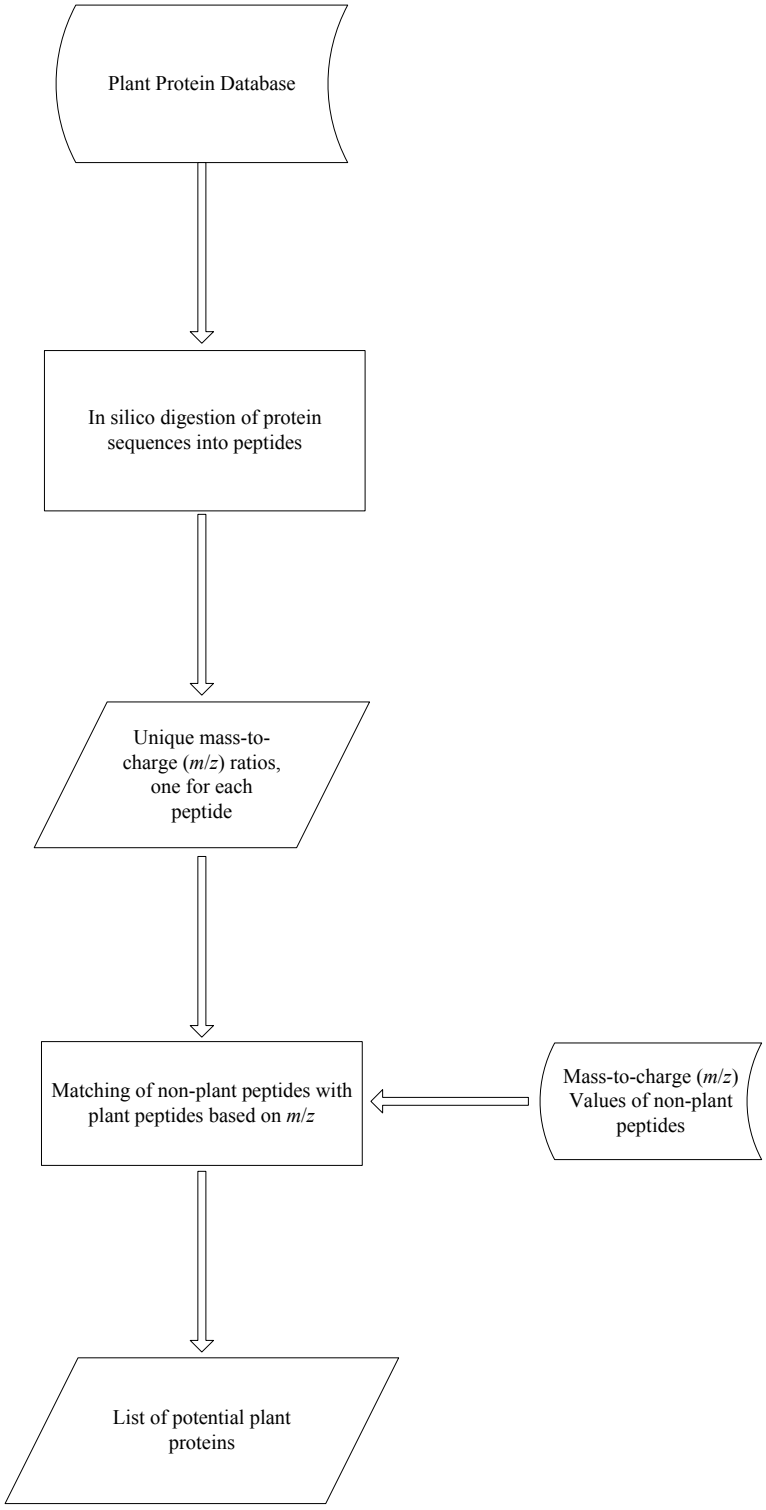

Figure 3.1: Methodology to generate a list of potential plant proteins. The plant protein database is *in silico* digested into peptides using a residue-specific protease. This produces unique sets of peptides per protein. Each peptide is assumed to carry a unique m/z value and charge. The m/z values of non-plant peptides are correlated with the m/z values of plant peptides in the database. Based on the matches, a list of plant proteins is provided.

The non-plant peptides reported by Pro Group serve as input for the pipeline. Figure A.2 in Appendix A displays a portion of a Pro Group report showing accession numbers of the identified proteins, protein names and peptide sequences. The pipeline calculates the most likely m/z value for each non-plant peptide. Calculation of m/z ratios involves determination of intermediate values such as basic AA count and MW. The m/z values are pre-calculated and associated with peptide entries in the processed plant protein database. The pipeline also requires a threshold or tolerance value. The units of this value are Parts Per Million (PPM). Based on this tolerance, the pipeline correlates m/z values of non-plant peptides with the m/z values of plant peptides. The output is a list of proteins. The list consists of protein, score pairs. The score reflects total number of peptides which were found associated with the non-plant m/z value.

There are various components to the bioinformatics pipeline described in this thesis. Earlier, Figure 3.1 showed the implemented methodology in a generalized manner. Figure 3.2 shows various specific stages involved in the plant protein identification/recovery. Details of various steps are provided in this section. The first step involves collection of data and processing of the reference data sets. Once collected, each peptide (in the database and input non-plant set) has its MW, m/z value, charge state and count of basic AAs calculated. The m/z value associated with the non-plant peptide is then correlated with the m/z values of the peptides present in the plant database. The pipeline then produces three lists of potential plant proteins. The first potential plant list is solely based on m/z values. The second potential list compensates for the multiple occurrences of identical peptides associated with the identified proteins in the first list. The third list corrects the bias of the pipeline towards longer (heavier) proteins. Intuition suggests that proteins present consistently across the three reported lists have greater chances of being present in the original sample.

3.2.1 Data Sources and Data Processing

The following discussion concerns steps 1(b), 2(b) and 3(b) of Figure 3.2. Most of the computer algorithms in the pipeline were implemented using Perl (version 5.8.8). Awk and Bash scripts were also used. The plant database consists of NCBI-REFSEQ restricted to green plants. It was obtained in FASTA format. The database was downloaded on 2007-10-01 and contained 11013 protein sequences. Tools from the EMBOSS [35] software package such as digest, seqret and supermatcher were also used. Unless otherwise stated, most of the processing was performed on an IBM computer with a dual Intel 1.4GHz CPU, 3GB of RAM, and the Linux operating system. A total of thirteen experimental reports were available for analysis. Three Pro Group reports were selected at random and non-plant peptides from each of the reports were collected manually and stored in three separate text files. All the non-plant peptides from each of these reports were used as input in three separate experiments (more details are provided in Chapter 4).

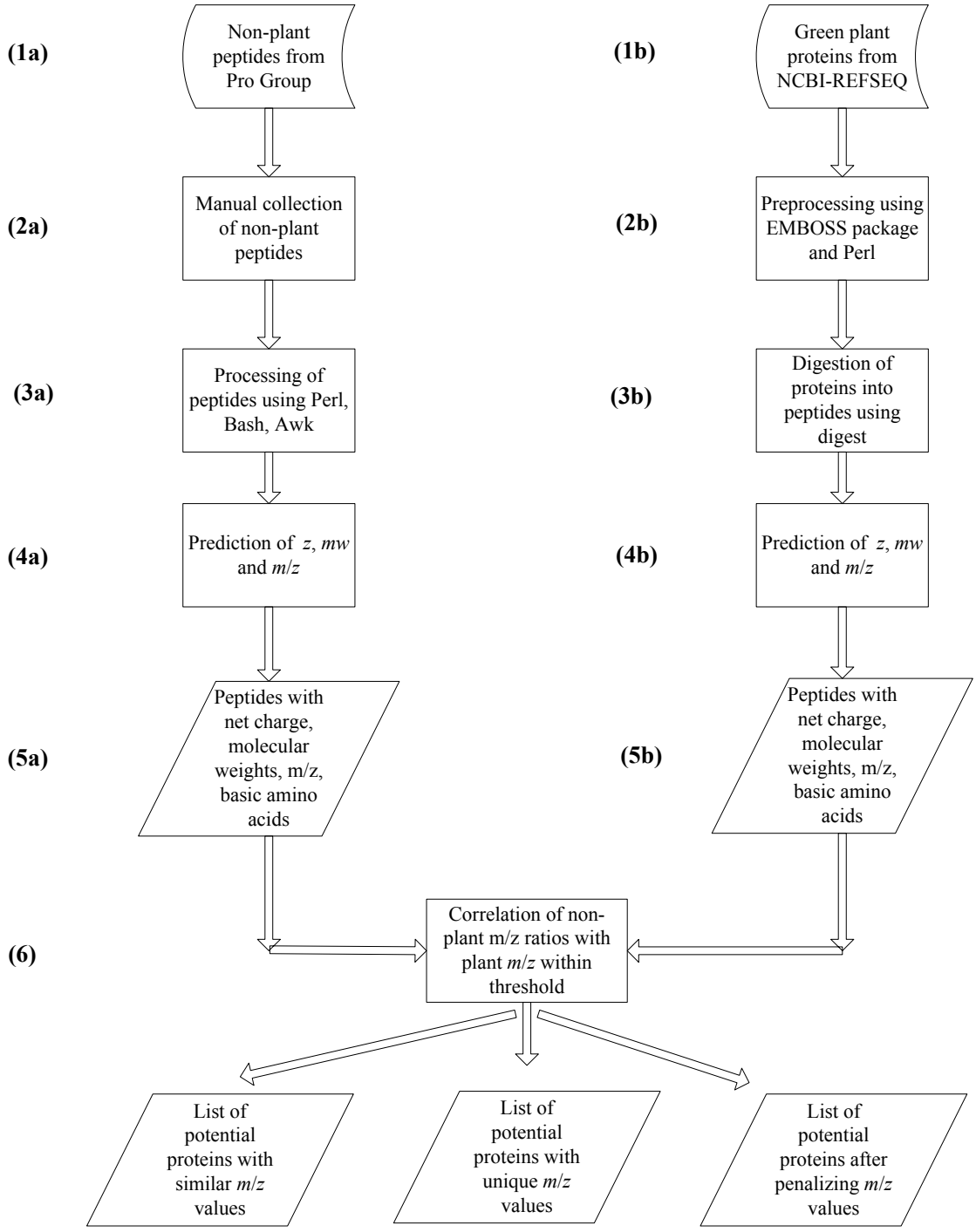

Figure 3.2: Stages in the pipeline. The figure shows various steps involved in the pipeline presented in this thesis. Non-plant peptides from Pro Group (1a) and plant peptides from NCBI-REFSEQ (1b) were the two sources of data. Various processing steps were needed (2a, 2b) and (3a, 3b) before m/z, MW, and charge could be predicted for non-plant and plant peptides (4a, 4b). After processing each peptide was assigned m/z, MWs and charges (5a, 5b). Subsequently, m/z ratios of non-plant peptides were correlated with m/z of plant peptides based on a threshold value set by the user (6). Three lists of potential plant proteins were the output.

3.2.1.1 Database Processing

The database used for the protein identification was restricted to green plants (Viridiplantae). Use of restricted databases can be a powerful mechanism for reducing false positives. It should eliminate matches from other species. However, there is a drawback to this approach. If there is a novel protein present in the sample – one not present in the database – then it will not be identified, i.e. there is a greater risk of false negatives.

The complete plant protein database was processed to remove redundant protein sequences. For example, identical entries of protein sequences were found:

>gi|118430281|ref|YP_874715.1| ribosomal protein S19 [Agrostis stolonifera]
MTRKKTNPFVAHHLLAKIEKVNRKEEKETIVTWSRASSILPTMVGHTIAIHNGKEHIPIYITNPMVGRKLGEFVPTRHF
TSYENARKDTKSRR
>gi|118430342|ref|YP_874777.1| ribosomal protein S19 [Agrostis stolonifera]
MTRKKTNPFVAHHLLAKIEKVNRKEEKETIVTWSRASSILPTMVGHTIAIHNGKEHIPIYITNPMVGRKLGEFVPTRHF
TSYENARKDTKSRR

In such a case, only one sequence was retained. Fragments of proteins also in the database were removed as well. The processing was performed using a Perl script written by a fellow researcher, Brett Trost (unpublished). The scanning and processing were completed in a wall-clock time of under 24 hours. This was acceptable since the operation only needed to be performed once.

The use of an alternate plant database from TAIR (The Arabidopsis Information Resource)[1] was also investigated. It was found that the TAIR database contained redundant protein sequences. Furthermore, sequences from non-plant species, such as viruses were also present. For example, the following entry was found in the database:

>gi|3184156|emb|CAA04392.1|ORFA+B[Vicia faba] endornavirus.

Therefore, further use of TAIR database was abandoned.

The pipeline uses m/z ratios of peptides for identification. All the protein sequences were digested *in silico* into peptides using the digest program from EMBOSS. The program accepts protein sequences in FASTA format as input. By default, digest uses the average mass of AAs. In order to have digest use the monoisotopic mass of the 20 AAs, the entries in the EMBOSS data file *Eamino.dat* were changed to reflect the monoisotopic mass of amino acids. The enzyme trypsin was selected to digest the protein sequences in the database. Figure 3.3 shows a partial example of a digestion of a protein into peptides with monoisotopic mass.

It is generally accepted that trypsin cleaves a protein at an R or K not followed by P at the C-terminus. However, at a more detailed level there are variations to this notion. Under digest, for instance, trypsin will not favor a cut site with K or R if it is followed by any of KRIFLP [1]. For

[1] http://www.arabidopsis.org/

>YP_874716.1 YP_874716.1 photosystem II protein D1 [Agrostis stolonifera]
MTAILERRESTSLWGRFCNWITSTENRLYIGWFGVLMIPTLLTATSVFIIAFIAAPPVDIDGIREPVSGSLLYGNN
IISGAIIPTSAAIGLHFYPIWEAASVDEWLYNGGPYELIVLHFLLGVACYMGREWELSFRLGMRPWIAVAYSAP
VAAATAVFLIYPIGQGSFSDGMPLGISGTFNFMIVFQAEHNILMHPFHMLGVAGVFGGSLFSAMHGSLVTSSLI
RETTENESANEGYKFGQEEETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVVGIWFTALGISTMAFNL
NGFNFNQSVVDSQGRVINTWADIINRANLGMEVMHERNAHNFPLDLAALEVPSING

Mol_Weight	cterm	nterm	Sequence
10328.205	R	E	EWELSFRLGMRPWIAVAYSAPVAAATAVFLIYPIGQGSFSDGMPLGISGTFN FMIVFQAEHNILMHPFHMLGVAGVFGGSLFSAMHGSLVTSSLIR
7036.562	R	E	EPVSGSLLYGNNIISGAIIPTSAAIGLHFYPIWEAASVDEWLYNGGPYELIVLH FLLGVACYMGR
6271.266	R	E	ESTSLWGRFCNWITSTENRLYIGWFGVLMIPTLLTATSVFIIAFIAAPPVDIDGIR
5080.316	R	S	ETTENESANEGYKFGQEEETYNIVAAHGYFGRLIFQYASFNNSR
4758.383	R	V	SLHFFLAAWPVVGIWFTALGISTMAFNLNGFNFNQSVVDSQGR
1991.010	R	.	NAHNFPLDLAALEVPSING
1313.709	R	A	VINTWADIINR
1285.590	R	N	ANLGMEVMHER
988.548	.	E	MTAILERR

Figure 3.3: Peptide sequences with monoisotopic mass after using the digest program from EMBOSS. Figure shows a portion of the output from the digest program for the protein sequence YP_874716.1 from the NCBI-REFSEQ database. The monoisotopic MW is given under Mol_Weight. The values in this column utilize the monoisotopic weights of the AAs present in the sequence (including the monoisotopic weight of water).

example, for the following entry in the database digest produced no fragments (tryptic peptides):

>gi|118430285|ref|YP_874719.1| photosystem II protein K [Agrostistolonifera]

MPNILSLTCICFNSVLYPTTSFFFAKLPEAYAIFNPIVDVMPVIPLFFFLLAFVWQAAVSFR

If the general rule for tryptic digestion is followed, two fragments should have been produced: MPNILSLTCICFNSVLYPTTSFFFAK and LPEAYAIFNPIVDVMPVIPLFFFLLAFVWQAAVSFR

The digest program classifies cleavage sites as favorable or unfavorable. For example, setting the unfavorable option would produce all the fragments ending with K followed by any of KRIFLP, or R followed by any of KRIFLP. A user can control the operation of digest by specifying whether unfavorable cleavage sites are produced, or only favorable ones. In the above example the favorable option was selected. Selection of unfavorable cleavage sites would have resulted in an increase in the number of digestion products in the results. Furthermore, this would also have increased the time required for data processing by increasing the size of the search space. Hence a favorable digestion was specified.

Another program, PeptideCutter from ExPASy, was also investigated. This program follows a slightly different set of rules governing where proteins are cleaved when trypsin is selected as the proteolytic enzyme. The set of rules is explained on the ExPASy website[2]. The paper by Gasteige et al. [16] also provides a brief overview of PeptideCutter. PeptideCutter was not used for processing since it lacks an interface usable in batch mode or capable of uploading a whole database for

[2]http://ca.expasy.org/tools/peptidecutter/peptidecutter_enzymes.html. Website last accessed on August 16, 2008.

processing. As a result, the digest program from EMBOSS package was used for the digestion of the reference database.

Debate regarding the specific products of trypsin digestion still continues in the bioinformatics community [36]. Figures 2.3 and 2.9 demonstrate that apart from mis-cleavage sites, variability in the knowledge, understanding and behavior of proteolytic enzymes affects the correct identification of peptides.

Protein sequences from the standard databases sometimes contain characters that do not correspond to the 20 known AAs. Characters such as B, X, U, Z are sometimes present. For example, the following peptide sequence with said characters was found in the reference database:

MATNGNGASGAELATFALGUFWHPEASFANVPGVVKXIIXMHKPBPFXISCNK.

Any digest products/peptides containing such characters were removed from further analysis. This removal creates artifactual sequences; it results in retaining only a part of the original sequence. However, this strategy was by far the best for information conservation purposes. The alternative would be to remove the entire protein sequence containing such non-standard residue designations. The latter strategy would lead to excessive information loss.

The program digest provided much information which is not used in the pipeline. Hence, the output file from digest was filtered to only include necessary details. Figure 3.4 shows a sample of the original and the modified output from the program digest. Information such as sequence length and fragment counts (number of tryptic digests) was removed. Further, Pro Group does not consider peptides which are less than 400 Da and more than 6000 Da. Therefore, all the peptides whose molecular weights were not between these two values were removed from the digest output. This one-time preprocessing of the information from the plant-only database was completed in a reasonable amount of wall-clock time (under 24 hours).

(1)

>gi|118430282|ref|YP_874716.1| photosystem II protein D1 [Agrostis stolonifera]
MTAILERRESTSLWGRFCNWITSTENRLYIGWFGVLMIPTLLTATSVFIIAFIAAPPVDIDGIREPVSGSLLYGNNIIS
GAIIPTSAAIGLHFYPIWEAASVDEWLYNGGPYELIVLHFLLGVACYMGREWELSFRLGMRPWIAVAYSAPVAAAT
AVFLIYPIGQGSFSDGMPLGISGTFNFMIVFQAEHNILMHPFHMLGVAGVFGGSLFSAMHGSLVTSSLIRETTENES
ANEGYKFGQEEETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVVGIWFTALGISTMAFNLNGFNFNQSVV
DSQGRVINTWADIINRANLGMEVMHERNAHNFPLDLAALEVPSING

(2)

```
#=======================================
#
# Sequence: YP_874716.1    from: 1   to: 353
# HitCount: 9
#
# Complete digestion with Trypsin yields 9 fragments
#
#=======================================
```

Start	End	Mol_Weight	cterm	nterm	Sequence
130	225	10328.205	R	E	EWELSFRLGMRPWIAVAYSAPVAAATAVFLIYPIGQGSFSDGM PLGISGTFNFMIVFQAEHNILMHPFHMLGVAGVFGGSLFSAM HGSLVTSSLIR
65	129	7036.562	R	E	EPVSGSLLYGNNIISGAIIPTSAAIGLHFYPIWEAASVDEWLYNG GPYELI VLHFLLGVACYMGR
9	64	6271.266	R	E	ESTSLWGRFCNWITSTENRLYIGWFGVLMIPT LLTATSVFIIA FIAAPPVDIDGIR
226	269	5080.316	R	S	ETTENESANEGYKFGQEEETYNIVAAHGYFGRLIFQYASFNNSR
270	312	4758.383	R	V	SLHFFLAAWPVVGIWFTALGISTMAFNLNGFNFNQSVVDSQGR
335	353	1991.010	R	.	NAHNFPLDLAALEVPSING
313	323	1313.709	R	A	VINTWADIINR
324	334	1285.590	R	N	ANLGMEVMHER
1	8	988.548	.	E	MTAILERR

```
#---------------------------------------
#---------------------------------------
```

(3)

Protein: YP_874716.1

Start	End	Mol_Weight	cterm	nterm	Sequence
130	225	10328.205	R	E	EWELSFRLGMRPWIAVAYSAPVAAATAVFLIYPIGQGSFSDGM PLGISGTFNFMIVFQAEHNILMHPFHMLGVAGVFGGSLFSAM HGSLVTSSLIR
65	129	7036.562	R	E	EPVSGSLLYGNNIISGAIIPTSAAIGLHFYPIWEAASVDEWLYNG GPYELI VLHFLLGVACYMGR
9	64	6271.266	R	E	ESTSLWGRFCNWITSTENRLYIGWFGVLMIPT LLTATSVFIIA FIAAPPVDIDGIR
226	269	5080.316	R	S	ETTENESANEGYKFGQEEETYNIVAAHGYFGRLIFQYASFNNSR
270	312	4758.383	R	V	SLHFFLAAWPVVGIWFTALGISTMAFNLNGFNFNQSVVDSQGR
335	353	1991.010	R	.	NAHNFPLDLAALEVPSING
313	323	1313.709	R	A	VINTWADIINR
324	334	1285.590	R	N	ANLGMEVMHER
1	8	988.548	.	E	MTAILER

Figure 3.4: Selective information extracted from the output of program digest. Each protein sequence in the database (1) was digested into peptides (2). The file was processed to retain only a subset of the information (3).

3.2.1.2 Input Data Processing

The following discussion concerns steps (1a), (2a), (3a) of Figure 3.2. Non-plant peptides reported by Pro Group were the inputs to the pipeline. Pro Group output is in the form of a Microsoft Excel spreadsheet (see Figure 2.1). Selecting a given protein in that spreadsheet automatically highlights the peptides associated with it. Using this interface, all the non-plant peptides were selected manually. For testing purposes plant peptides, non-plant peptides and a combination of plant and non-plant peptides were used as input. Figure 3.5 shows a partial list of non-plant peptides reported by Pro Group. As an example, these non-plant peptides came from the following organisms (the names present here are taken verbatim from the Pro Group report): Pig, *Thermosynechococcus elongatus*, *Homo sapiens*, Mouse, Sea star, *Mycobacterium tuberculosis*, *Calloselasma rhodostoma*, and *Plasmodium falciparum*.

The input list of non-plant peptides was preprocessed using a Perl script before being used in the pipeline for plant protein identification. Pro Group reports the AA K as J, and Y as U. The Perl Script replaced all of the J's with K's, and U's with Y's, in the input sequence data.

IITHPNFNGNTLDNDIMLIJ
LGEHNIDVLEGNEQFINAAJ
LSSPATLNSR
VATVSLPR
LAFYDYIGNNPAJ
VHTVLINDPGR
AJPVSSGSPWYGSDR
AKPVSSGSPWYGSDR
ELEVIHCR
IGGIGTVPVGR
QTVAVGVIJ
AQYEDIAQJ
SEIDNVJ

Figure 3.5: Partial list of non-plant peptides used as input sequences. These non-plant peptides were reported by Pro Group and were manually selected for processing and analysis.

3.3 Prediction of Molecular Weights, Charge State, and Mass-to-Charge Ratios

One of the very important parts of the bioinformatics pipeline described in this thesis is the prediction of MWs, charge state(s), and mass-to-charge ratio(s) of peptides.

In MS/MS, mass (m) and charge (z) of the peptides are the two important properties for protein identification. Peptides (daughter-ions) are composed of chains of AAs. Each peptide ion has a unique mass and charge associated with it. Mass refers to the combined monoisotopic, residual mass of each AA present in a peptide. For example, peptide MTAILERR has a mass of 988.5484:

M+T+A+I+L+E+R+R+H_2O (residual, monoisotopic mass of each AA + monoisotopic mass of water) or 131.04049+101.04768+71.03711+113.08406+113.08406+129.04259+156.1011+156.1011+18.01056. The weight of water is added because of flanking H and OH ions at N and C termini, respectively.

Every peptide sequence carries a net charge. The peptides can be either negatively or positively charged. The net charge of a tryptic peptide depends on the charge at the N-terminus and the C-terminus, and the number of basic AAs H (Histidine), K, and R present. In ESI, most peptide ions in the solution are protonated (have a proton, H^+, added) and processed in a positive ion mode, and thus carry a net positive charge [25]. Peptides from ESI-MS/MS carry multiple charges, of which doubly (2^+) charged ions are dominant. Further, singly (1^+) and triply (3^+) charged ions are also frequently observed in a mass spectrum [14, 25]. The raw peptides (from Pro Group) used as input to the pipeline should have at least a charge of 2^+ and no charge greater than 4^+.

3.3.1 Predictions for Peptides in the Reference Database

The following discussion relates to steps (4b) and (5b) of Figure 3.2. A Perl script from the bioinformatics pipeline was used to calculate the MW of all the tryptic peptides present in the reference database, and to predict their charge (z) state(s) and m/z ratios. For each protein, the MWs of its tryptic peptides were calculated by summing the residual monoisotopic mass of each AA present in the peptide and adding the monoisotopic mass of the water.

Since the actual mass spectrum based on which Pro Group reported plant and non-plant proteins was not available, it was difficult to accurately predict m/z ratios and the charge states of the peptides. Because the ESI-MS/MS approach was used, peptides would have multiple charges in the mass spectrum. In order to predict the charge state of peptides produced by ESI, the following approach was adopted: The (flanking) C-termini of tryptic peptides were taken to be always positively charged (+1). Assuming the (flanking) N-termini always bear a positive charge (+1), the net charge (total charge) of the peptide was calculated by adding one positive charge (+1) for each occurrence of a basic AA in the peptide sequence.

For each peptide in the database, possible charge states were calculated (Table 3.1) and for each such peptide the m/z ratio was calculated as: $(MW+(charge*1.0072))/charge$, where MW is the molecular weight of the peptide, $charge$ is the predicted charge, and 1.0072 is mass of one proton. Figure 3.6 displays the calculation results for one protein from the database. In the figure, the MW of MTAILERR is 988.5484, and its predicted charges are 2^+, 3^+ and 4^+. For z=2^+, the m/z is (988.5484+(2*1.0072))/2 = 495.2814. The other m/z values of 330.5233 and 248.1433 were obtained similarly.

The Perl scripts used in the pipeline for calculation of MW and m/z for the input data and m/z values for peptides in the reference database were set to use values with four significant figures.

The digest program calculated the MW to three significant figures. While processing the output from digest, the MW of peptides in the databases were recalculated to four significant figures to avoid any roundoff errors.

Table 3.1: Assumed charges on the peptides in the reference database

Number of Basic Amino Acids	Charges
1	3^+
2	2^+, 3^+
3	2^+, 3^+, 4^+
4	3^+, 4^+
>4	4^+

Start	End	Mol_Weight	m/z	Charge	BasicAA	ProteinID	Sequence
335	353	1991.0105	664.6774	3	2	YP_874716.1	NAHNFPLDLAALEVPSIN
335	353	1991.0105	996.5124	2	2	YP_874716.1	NAHNFPLDLAALEVPSING
313	323	1313.7088	438.9101	3	2	YP_874716.1	VINTWADIINR
313	323	1313.7088	657.8616	2	2	YP_874716.1	VINTWADIINR
324	334	1285.5902	322.4047	4	3	YP_874716.1	ANLGMEVMHER
324	334	1285.5902	429.5373	3	3	YP_874716.1	ANLGMEVMHER
324	334	1285.5902	643.8023	2	3	YP_874716.1	ANLGMEVMHER
1	8	988.5484	248.1443	4	3	YP_874716.1	MTAILERR
1	8	988.5484	330.5233	3	3	YP_874716.1	MTAILERR
1	8	988.5484	495.2814	2	3	YP_874716.1	MTAILERR

Figure 3.6: Entries in the processed database. For each peptide, multiple charges were calculated based on the number of basic AAs. The heading BasicAA refers to the total count of H and R. Predicted charges based on the number of basic AA are shown under Charge. Molecular weight shown under the heading Mol_weight refers to the sum of residual, monoisotopic masses of each AA in the peptide. The m/z ratio of each peptide is under m/z. The values under Start and End refer to the position (from 1 to n, where n is the length of the peptide) at which the enzyme cleaved the protein into a unique peptide. Initially these latter values were used while performing manual validation of the pipeline. They have no significance as such in the algorithm. The figure only shows first few lines of the entries.

For faster protein identification all the peptide sequences in the database were sorted in increasing order based on m/z value of the peptides. Figure 3.7 shows a sample of the final output.

Start	End	Mol_Weight	m/z	Charge	Basic AA	ProteinID	Sequence
877	879	430.2328	108.5654	4	3	XP_001420882.1	HFK
1020	1022	430.2651	108.5735	4	3	XP_001420661.1	QKR
102	104	430.2651	108.5735	4	3	XP_001419999.1	QRK
1044	1046	430.2651	108.5735	4	3	XP_001421378.1	QKR
1045	1047	430.2651	108.5735	4	3	XP_001417480.1	QKR
1085	1087	430.2651	108.5735	4	3	YP_636457.1	QKR
109	112	430.2651	108.5735	4	3	XP_001418837.1	AGRK
1103	1105	430.2651	108.5735	4	3	XP_001417492.1	QKR
1104	1106	430.2651	108.5735	4	3	XP_001416340.1	QKR
111	113	430.2651	108.5735	4	3	YP_717292.1	QRK
112	114	430.2651	108.5735	4	3	XP_001419830.1	QRK
117	119	430.2651	108.5735	4	3	XP_001422727.1	QKR
118	121	430.2651	108.5735	4	3	XP_001418035.1	AGKR
119	121	430.2651	108.5735	4	3	XP_001415387.1	QKR
1199	1201	430.2651	108.5735	4	3	XP_001421977.1	QKR
124	126	430.2651	108.5735	4	3	XP_001417464.1	QRK
130	132	430.2651	108.5735	4	3	XP_001421977.1	QKR
1313	1315	430.2651	108.5735	4	3	XP_001417100.1	QKR
1327	1329	430.2651	108.5735	4	3	XP_001417100.1	QKR
1	3	430.2651	108.5735	4	3	XP_001416854.1	QKR
140	142	430.2651	108.5735	4	3	XP_001420072.1	QRK
143	146	430.2651	108.5735	4	3	XP_001419698.1	GAKR

Figure 3.7: Final output from the pipeline after processing the plant database. First, all the proteins were digested using digest from EMBOSS. Peptides containing ambiguous AAs were removed. For each peptide MW, m/z, charge, and the number of basic AAs were calculated. The file was then sorted in increasing order of m/z values.

3.3.2 Predictions for Non-Plant Peptides

The following discussion reflects steps (4a) and (5a) of Figure 3.2. The monoisotopic MW, charge state(s) and m/z value(s) of the input data (peptides) were necessary for the pipeline. Recall the method described earlier for calculating the MWs of the plant peptides. Since they were not available in the Pro Group output, they were predicted based on the peptide fragments. The same method was followed to calculate MWs of non-plant peptides. The only modification was that instead of calculating multiple charges for each non-plant peptide, only the highest possible charge was considered. Table 3.2 lists how charges were predicted for non-plant peptides.

In an earlier version of the pipeline, multiple charges were considered for non-plant peptides as described in Table 3.1. However, during analysis of the results obtained after testing the pipeline, it was observed that erroneous peptides were reported (details provided in the next section). Hence, the single most intense charge was used.

Table 3.2: Assumed charge on the non-plant peptides

Number of Basic Amino Acids	Charges
2	3^+
3	4^+
4	4^+
>4	4^+

3.4 Algorithm

This section describes how m/z values for non-plant peptides (input data) and plant peptides (protein database) were used to generate a list of potential plant proteins.

3.4.1 Correlation of Input Data with Reference Database

The discussion presented in this section concerns step 6 of Figure 3.2. The identification of proteins from MS spectra can be described in general terms as follows: by using a site-specific protease, proteins are *in silico* digested into peptides that carry unique m/z ratios. These ions are correlated with digested proteins from the sequence database for protein identification. A certain number of measured/observed peptides have to agree with the predicted peptides from the database before a protein is considered as identified. The number of peptide matches contributes to the confidence level for the identified protein. This does not mean that a single peptide match should not be considered as having identified a protein. However, we would expect that cases where a peptide was identified by only a single match would be rare. It is an accepted fact that the experimental m/z values will almost always be slightly different from the actual m/z of the peptides

being analyzed due to errors such as instrument errors, contaminations, PTMs and miss-cleavage sites. In order to compensate for such differences between measured and predicted m/z ratios the search of the sequence database should be error tolerant. Therefore, the protein identification software tries to compensate for such errors by providing a user-settable threshold value (error tolerance).

In the bioinformatics pipeline presented in this thesis, a threshold value in PPM (Parts Per Million) is used to compensate for such errors. For each m/z value of a non-plant peptide in the input, an interval with an upper limit (U_limit) and a lower limit (L_limit) is determined. The U_limit and L_limit are calculated based on the error tolerance as shown in Figure 3.8. The pipeline correlates each non-plant m/z value with the theoretical m/z values of the peptides from the plant database and retains all the peptides which are within the specified interval. This is repeated for every m/z value in the input non-plant set.

Because the peptides from the database are sorted by the m/z ratio, the scanning of plant peptides can be terminated as soon as an m/z value greater that than the U_limit is encountered. Figure 3.8 shows how the Perl script scans the file, identifies the peptides within the set interval and retains them. Figure 3.9 shows intermediate results when a threshold value of 100 PPM (0.0001) is specified (Figure 3.10 shows the raw data that was used as input). Details of the file hits.*var* will be described later in this section.

The logic behind this portion of the pipeline is explained with the help of an example. Suppose a non-plant peptide TGKRLTR was part of the input to the pipeline and that the user specified an error tolerance of 100 PPM. The highlighted row of Figure 3.9 shows the non-plant peptide sequence has an m/z of 208.6343. Based on the error tolerance an U_limit of 208.6551 and L_limit of 208.6134 were calculated (stage 2 of Figure 3.9). The plant-only database was scanned and all the peptides in the interval between the two limits were collected and stored in a file hits.208.6343 (stage 3 of Figure 3.9). In this pipeline, a large amount of information (m/z values and associated information, peptide sequences, MWs etc.) was generated and recorded.

The information is stored in intermediate files with names of the form hits.*var* (stage 4 of Figure 3.9). Here, *var* represents the m/z value of an input non-plant peptide. A hits.*var* file is created for each unique m/z ratio of an input non-plant peptide. In each of these files, all the peptides within the set range are collected. In Figure 3.9 there are 20 non-plant peptides, each with a m/z ratio. Hence there are 20 hits.*var* files, each containing correlated plant peptides. From these files, information such as protein name, m/z ratios, and MW can be used for testing and verification purposes. Testing is an important component of any software design. A complete description of the pipeline's testing is discussed later in Section 3.5.

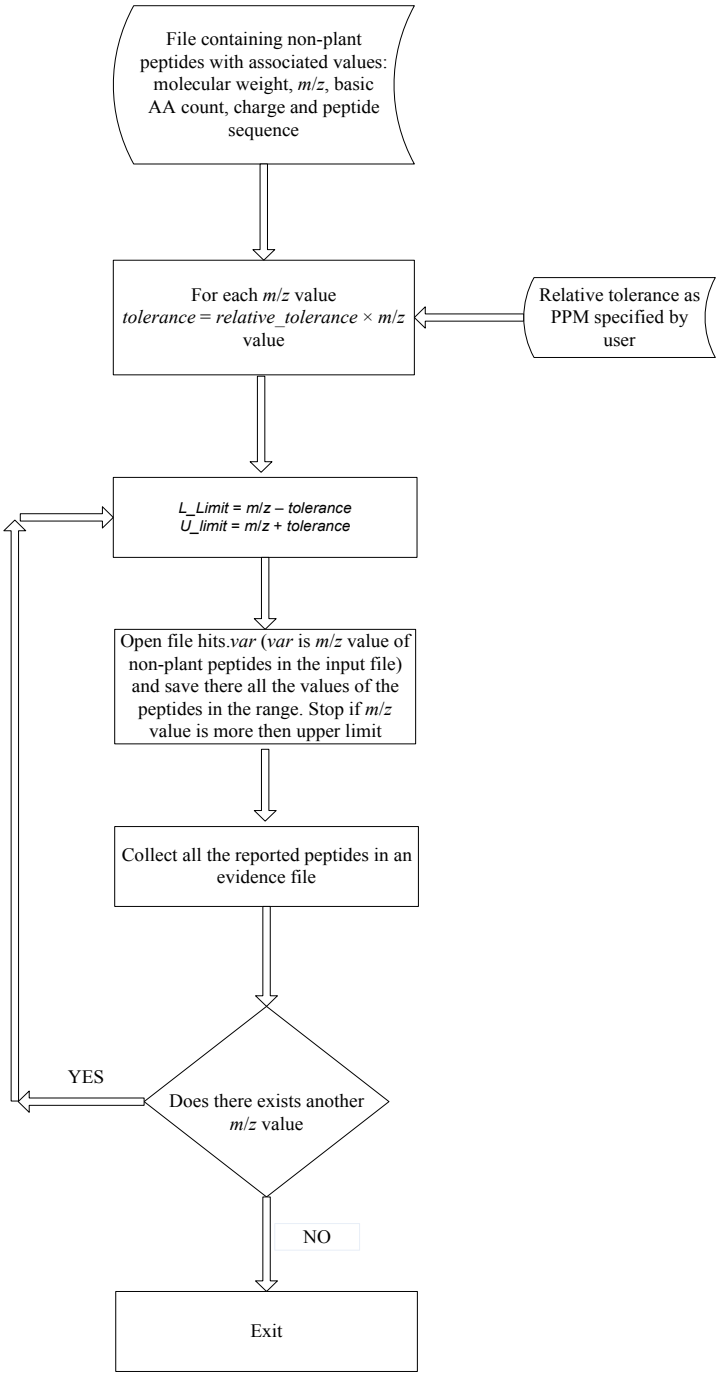

Figure 3.8: Flowchart showing how non-plant peptides are correlated with plant peptides based on their m/z values by the Perl script in the pipeline. The script captures the m/z value associated with each non-plant peptide from the input file. For each value, an upper and lower limit (an interval) is calculated. The two limits depend on the relative threshold value entered by the user. Once the two limits are calculated, the script scans the plant database and captures all the peptides within the set interval. The script will not scan further if the m/z value in the database is more than the upper limit. All the peptides within the range are retained in a separate text file. There is one file for each unique m/z value in the non-plant input. These files are important as they are the evidence upon which the pipeline reports potential plant proteins. The files also also used for testing/analyzing the pipeline.

(1)

Mol_Weight	m/z	BasicAA	Charge	Sequence
830.5083	208.6343	4	4	TGKRLTR
1218.6352	305.6660	4	4	DKRPEGYNLK
1014.6068	339.2095	3	2	SLLGNLGITK
1385.8277	347.4641	4	4	AKLVPKFLEDVK
1490.6242	373.6632	4	3	HFPWADGTSCGEGK
1550.7360	388.6912	4	3	DAEHYDTAILFTR
1193.6554	398.8923	3	2	AFFALVTNGVR
1614.8457	404.7186	4	3	GSQGKIVDLVDELDK
1671.9699	418.9997	4	5	VSDALLEKKKLMAAR
1800.9880	451.2542	4	3	TTTPVYVALGIFVQHR
1421.5544	474.8587	3	2	SCNIEDCPENNGK
1458.7137	487.2451	3	2	ASFGSGPPVEWTPK
2104.1131	527.0355	4	3	QSFVGMLTITDFINILHR
1746.9258	583.3158	3	2	QSNTSNIFLSPVTIAR
1885.9414	629.6543	3	2	APDLPSESGSPVYVNQVK
2008.0047	670.3421	3	2	WLPQQNAAYFLLSTNDK
2009.1398	670.7205	3	2	GIVSLSDILQALVLTGGEKP
2689.3062	673.3337	4	4	VWDLNMENRPIETYQVHNYLR
2020.0064	674.3427	3	2	AQLITDSPGSTSSVTSINSR
2918.4473	730.6190	4	4	SFSLKLSSSISDVKFSQQWEDIMTR

(2)

var	L_limit	U_limit
208.6343	208.61343657	208.65516343

(3)

Mol_Weight	m/z	BasicAA	Charge	ProteinID	Sequence
830.4284	208.6143	4	3	XP_001416521.1	HAYEIAK
830.4283	208.6143	4	3	XP_001420843.1	HLGFTEK
415.2178	208.6161	2	2	YP_001004215.2	SGPR
415.2178	208.6161	2	2	YP_001123228.1	SGPR
415.2178	208.6161	2	2	YP_001123491.1	SGPR
415.2178	208.6161	2	2	YP_001123579.1	SGPR
.					
.					
415.2794	208.6469	2	2	XP_001420636.1	VAVK
415.2794	208.6469	2	2	XP_001419027.1	VVAK
415.2794	208.6469	2	2	XP_001417143.1	AVVK
415.2794	208.6469	2	2	XP_001418250.1	VVAK
415.2794	208.6469	2	2	XP_001417706.1	AVVK
415.2906	208.6525	2	3	XP_001415730.1	KLR

(4)

hits.208.6343	hits.373.6632	hits.418.9997	hits.527.0355	hits.670.7205
hits.305.6660	hits.388.6912	hits.451.2542	hits.583.3158	hits.673.3337
hits.339.2095	hits.398.8923	hits.474.8587	hits.629.6543	hits.674.3427
hits.347.4641	hits.404.7186	hits.487.2451	hits.670.3421	hits.730.6190

Figure 3.9: Use of upper limit, lower limit and evidence file by the pipeline. Information such as MW, m/z ratio, charge, count of basic AAs for each non-plant peptide was stored in tab-separated form (1). When a threshold value (100 PPM) was given to the Perl script, the script calculated an upper and a lower limit (2) and captured all the peptides falling in the interval based on the m/z value (3). The information from step 3 was stored in a file whose name is of the form hits.var. In this case, the file was named hits.208.6343. It is one of many files containing intermediate information generated by the pipeline (4).

VWDLNMENRPIETUQVHNULR
WLPQQNAAUFLLSTNDJ
DJRPEGUNLJ
AQLITDSPGSTSSVTSINSR
APDLPSESGSPVUVNQVJ
VSDALLEJJJLMAAR
TGJRLTR
QSNTSNIFLSPVTIAR
GSQGJIVDLVDELDJ
AJLVPJFLEDVJ
SLLGNLGITJ
QSFVGMLTITDFINILHR
GIVSLSDILQALVLTGGEJP
TTTPVUVALGIFVQHR
AFFALVTNGVR
DAEHUDTAILFTR
HFPWADGTSCGEGJ
ASFGSGPPVEWTPJ
SCNIEDCPENNGJ

Figure 3.10: Partial list of non-plant peptides used as input to the pipeline. In total 20 peptides were used as input.

After collecting all the plant peptides within the user-specified tolerance for each non-plant peptide, the script next identifies candidate proteins. The total count of peptides associated with the same protein is reflected by a score associated with that protein. A higher peptide count results in a higher score, reflecting a greater confidence in the reported protein. A list of potential plant proteins sorted by decreasing order of the score is provided to the user (Figure 3.11). This list is termed the initial list. Further, all the supporting evidence for each reported protein — i.e. peptides from the protein that were within searched m/z ranges — is stored in an evidence file by the Perl script (Figure 3.12). There is one evidence file per reported protein. Peptides given in the evidence file can be tracked back to hits.*var* files and from there to the original input non-plant peptide sequences. This initial list is referred to as the "list of potential plant proteins with similar m/z value" in Figure 3.2.

There are various aspects of the initial list that require explanation. Many identical peptides are reported in Figure 3.12. Some of these repeated peptide sequences are the result of repeats within the protein sequence; for example, the subsequence HAYLTGRR is repeated within the protein. Further, some of the peptides — though not identical — are composed of the same AAs and would have indistinguishable mass spectra. An example of this is the three peptides beginning with A, ending with R, and having L and G (in some order) in between. Further, in the filtered plant protein database, each peptide sequence can be present more than once, depending on the charge state (see Section 3.3). The information filtering using threshold m/z values is not designed to remove these types of repeated peptides. These repeats are problematic because they bias the initial list toward proteins containing such multiply-occurring peptides. Therefore, a second list of potential plant proteins was generated from the first set by only considering unique peptide occurrences.

```
Score Protein_ID
  10 XP_001416177.1
   7 XP_001415739.1
   7 XP_001416378.1
   7 XP_001416972.1
   6 XP_001421248.1
   6 YP_778511.1
   5 XP_001417211.1
   5 XP_001417790.1
   5 XP_001418473.1
   5 XP_001421657.1
   5 XP_001421977.1
   5 XP_001422630.1
   4 NP_683861.1
   4 XP_001416600.1
   4 XP_001416917.1
   4 XP_001417140.1
   4 XP_001417525.1
   4 XP_001418349.1
   4 XP_001419308.1
   4 XP_001420157.1
   4 XP_001420401.1
   4 XP_001421119.1
   4 XP_001421255.1
   4 XP_001421282.1
   4 YP_001001592.1
   4 YP_001001595.1
   4 YP_001294246.1
   4 YP_778553.1
```

Figure 3.11: Initial results from the pipeline showing scores and protein IDs. In the first entry of the figure the score of 10 means that the pipeline found 10 peptide sequences from the same protein which have a matching m/z ratio. The figure only shows first 28 lines from the output.

Evidence for protein XP_001416177.1:

1270	1273	415.2542	208.6343	2	2	XP_001416177.1	ALGR
2727	2730	415.2542	208.6343	2	2	XP_001416177.1	ALGR
794	797	415.2542	208.6343	2	2	XP_001416177.1	AGLR
1001	1008	972.5251	487.2697	2	4	XP_001416177.1	HAYLTGRR
2458	2465	972.5251	487.2697	2	4	XP_001416177.1	HAYLTGRR
2078	2097	2104.0338	527.0156	4	3	XP_001416177.1	GPAMPALDVQPCVHGASWLR
4167	4186	2104.0338	527.0156	4	3	XP_001416177.1	GPAMPALDVQPCVHGASWLR
2046	2063	1746.9372	583.3196	3	3	XP_001416177.1	VSSSSAAAPFLAGALSRR
3422	3439	1746.9372	583.3196	3	3	XP_001416177.1	VSSSSAAAPFLAGALSRR
4135	4152	1746.9372	583.3196	3	3	XP_001416177.1	VSSSSAAAPFLAGALSRR

Figure 3.12: Contents of the evidence file for the highest scoring protein shown in Figure 3.11. All the peptides that could be produced from the identified protein that have an m/z ratio which matches a ratio from a non-plant peptide are listed. The 10 peptides give a cumulative score of 10 to the protein. From left-to-right, the following information is provided in the columns: Start, End, Molecular_Weight, m/z, Basic_AA, Charge, Sequence. For the meaning of these column headings refer to Figure 3.6.

This list is referred as the "list of potential plant proteins with unique m/z value" in Figure 3.2. The first occurrence of a peptide is retained and all others are removed (Figure 3.13). This typically results in the previously identified plant proteins receiving different scores. New evidence files are generated. The resultant output for our running example is shown in Figure 3.14 with sample evidence shown in Figure 3.15. It can be observed that the proteins appear in different orders in the two reported lists (Figures 3.11 and 3.14). Proteins reported in both lists have a greater chance of being present in the original sample.

Evidence for protein XP_001416177.1:

1270	1273	415.2542	208.6343	2	2	XP_001416177.1	ALGR
794	797	415.2542	208.6343	2	2	XP_001416177.1	AGLR
1001	1008	972.5251	487.2697	2	4	XP_001416177.1	HAYLTGRR
2078	2097	2104.0338	527.0156	4	3	XP_001416177.1	GPAMPALDVQPCVHGASWLR
2046	2063	1746.9372	583.3196	3	3	XP_001416177.1	VSSSSAAAPFLAGALSRR

Figure 3.13: Contents of evidence file containing only unique peptides. All the repeated, identical peptide sequences reported earlier (see Figure 3.12) for protein XP_001416177.1 are removed.

Another important observation can be made regarding the top proteins in the initial list. For example, in Figure 3.11 the top 5 proteins are XP_001416177.1, XP_001415739.1, XP_001416378.1, XP_001416972.1, and XP_001421248.1. These are long proteins; i.e, their sequence length is on the order of a few thousand to ten thousand AAs. Such heavy proteins were often observed in the top 5 positions on the reported initial plant list during testing of early versions of the pipeline. The report of such proteins is due to the fact that the most matching peptide fragments are reported by proteins with the most peptide fragments. A protein having more AAs results in an increased number of potential peptide fragments, and hence the probability of getting a match or hit to such a protein is higher than to a protein with a shorter sequence length. This observation is also reported

```
Score Protein_ID
7     XP_001416378.1
7     XP_001416972.1
6     XP_001421248.1
5     XP_001415739.1
5     XP_001416177.1
5     XP_001417211.1
5     XP_001418473.1
5     XP_001421657.1
5     XP_001421977.1
5     XP_001422630.1
4     NP_683861.1
4     XP_001416600.1
4     XP_001416917.1
4     XP_001417140.1
4     XP_001417525.1
4     XP_001418349.1
4     XP_001419308.1
4     XP_001420157.1
4     XP_001420401.1
4     XP_001421119.1
4     XP_001421255.1
4     XP_001421282.1
4     YP_001001592.1
4     YP_001001595.1
4     YP_001294246.1
4     YP_778553.1
3     NP_683769.1
3     NP_683776.1
```

Figure 3.14: Listing potential plant proteins with score based on only unique peptides. All the repeated, identical peptide sequences reported earlier (see Figure 3.12) are removed. The score for each protein is recalculated and a potential list is presented to the user. The figure only shows first 28 lines from the output.

by Berndt et al. [7]. This bias towards longer, heavier proteins is recognized by protein identification software (for example, MASCOT [33]) and is accordingly compensated for by various heuristic approaches. The goal is to penalize the score associated with heavy proteins. In our pipeline, to counterbalance the bias towards longer proteins, each protein score in the initial list is divided by an equilibrating factor, e. This factor is calculated for each protein in the database, and is based on the average length of fragments for proteins from the database. The following steps show how an equilibrating factor is calculated and used to generate a list of candidate proteins based on the initial list:

1) Find the total length of all proteins in the database.

2) Determine the count of all the peptide fragments present in the database.

3) Determine the average fragment length f from the total count of AAs in the database divided by total number of peptide fragments

4) Determine the length l for each protein reported in the initial list.

5) Calculate e for each protein as the ceiling of l/f.

Evidence for protein XP_001416378.1:

11520	11523	415.2542	208.6343	2	2	XP_001416378.1	GIAR
11318	11325	1014.5608	339.1941	3	2	XP_001416378.1	NTLWNLVR
17548	17558	1014.5706	339.1974	3	2	XP_001416378.1	GIAASVDAAIK
7180	7185	775.4451	388.7297	2	4	XP_001416378.1	TLIHHR
4311	4326	1746.8062	583.2759	3	2	XP_001416378.1	NSYLCYLAGMLSAEGR
5398	5410	1344.6628	673.3386	2	2	XP_001416378.1	VLEGSSAENLGNR
9787	9799	1346.7149	674.3646	2	2	XP_001416378.1	SLISSNTGNVSLR

Figure 3.15: Contents of the evidence file for the highest scoring protein in Figure 3.14. All the unique peptides that could be produced from the identified protein that have an m/z ratio which matches a ratio from non-plant peptides are listed. The 7 peptides give a cumulative score of 7 to the proteins. From left-to-right, the following information is provided in columns: Start, End, Molecular_Weight, m/z, Basic_AA, Charge, Sequence. For the meaning of these column headings refer to Figure 3.6.

6) Each protein present in the initial list has a score, s. Calculate an adjusted score, s' as s/e.

7) Re-sort the list according to the adjusted score, s'.

This list is referred to as the "list of potential plant proteins after penalizing the m/z value" in Figure 3.2. The list of proteins from this method for our running example is given in Figure 3.16.

Three different potential plant protein lists were the final output of the bioinformatics pipeline (Figure 3.17). Intuition would suggest that proteins consistently present across the three potential plant lists have a higher likelihood of being actually present in the sample. Hence in our running example, one would conclude that proteins XP_001420501.1 and XP_001421496.1 were present. Protein XP_001420501.1 is not among the top 30 hits in the second list in Figure 3.17. One would still consider this protein to potentially be in the sample, but with less confidence than the other two proteins.

There is not a set maximum to the number of proteins in the three lists reported by the pipeline.

For situations where no proteins are consistently reported across the three lists, a different threshold value should be selected and the pipeline re-run. A less stringent threshold should be selected as that will allow more proteins to be reported, and there will be a greater chance of some appearing consistently across multiple lists. In such cases user could investigate beyond the top 30 hits.

Score	Protein_ID
0.3333	YP_778511.1
0.2857	XP_001416438.1
0.2500	NP_683769.1
0.2500	NP_683794.1
0.2500	NP_817140.1
0.2500	NP_817210.1
0.2500	NP_817260.1
0.2500	NP_817262.1
0.2500	XP_001416183.1
0.2500	YP_001152065.1
0.2500	YP_001152090.1
0.2500	YP_001152098.1
0.2500	YP_001152160.1
0.2500	YP_636240.1
0.2500	YP_762308.1
0.2500	YP_784471.1
0.2000	NP_683849.1
0.2000	NP_817204.2
0.2000	XP_001415824.1
0.2000	XP_001417070.1
0.2000	XP_001417289.1

Figure 3.16: The list of potential plant proteins after compensating for longer proteins. Top proteins reported by the pipeline in Figure 3.11 were biased towards longer proteins. This fact is taken into account and the scores from the initial list are penalized by an equilibrating factor. The factor takes into account the average number of tryptic peptides per protein in the database. The scores were divided by this factor. The resultant list was ordered by the adjusted score. The figure shows the score and the associated protein ID. The figure only shows first 21 lines from the output.

Initial list based on m/z ratio			List based on unique peptides			List after correcting the bias for heavier proteins		
13	XP_001417215.1	1345	5	XP_001416048.1	4962	1.000	XP_001420501.1	52
6	XP_001416048.1	4962	4	YP_636192.1	2596	0.400	YP_001152205.1	58
6	XP_001418911.1	1823	3	XP_001416296.1	1280	0.375	XP_001421496.1	95
5	XP_001416378.1	18193	3	XP_001416378.1	18193	0.333	YP_001294304.1	37
5	XP_001418450.1	1701	3	XP_001417211.1	874	0.333	YP_636400.1	39
5	XP_001421646.1	560	3	XP_001417493.1	1135	0.285	NP_904239.1	81
4	XP_001417304.1	4003	3	XP_001418450.1	1701	0.285	XP_001419402.1	88
4	XP_001420501.1	52	3	XP_001418751.1	1418	0.250	YP_001152158.1	48
4	XP_001421699.1	645	3	XP_001418911.1	1823	0.230	XP_001418170.1	168
4	XP_001422420.1	1614	3	XP_001420239.1	3608	0.222	XP_001419514.1	115
4	YP_636192.1	2596	3	XP_001421496.1	95	0.222	XP_001419556.1	110
3	NP_084773.1	2280	3	XP_001421646.1	560	0.222	YP_025759.1	111
3	XP_001415435.1	223	3	XP_001422369.1	1091	0.200	NP_084760.1	54
3	XP_001416295.1	1051	3	XP_001422420.1	1614	0.200	XP_001418789.1	123
3	XP_001416296.1	1280	3	XP_001422699.1	861	0.200	XP_001419638.1	119
3	XP_001416870.1	1782	3	YP_001001595.1	2262	0.200	XP_001420788.1	126
3	XP_001417056.1	991	2	NP_084773.1	2280	0.200	XP_001420821.1	358
3	XP_001417211.1	874	2	XP_001415435.1	223	0.200	XP_001422464.1	121
3	XP_001417493.1	1135	2	XP_001415603.1	925	0.200	YP_025751.1	64
3	XP_001417762.1	611	2	XP_001415901.1	200	0.181	XP_001420067.1	134
3	XP_001417978.1	954	2	XP_001415907.1	365	0.181	XP_001420298.1	137
3	XP_001418147.1	596	2	XP_001416037.1	552	0.166	XP_001415435.1	223
3	XP_001418170.1	168	2	XP_001416081.1	664	0.166	XP_001415645.1	74
3	XP_001418654.1	293	2	XP_001416088.1	2253	0.166	XP_001416840.1	155
3	XP_001418751.1	1418	2	XP_001416284.1	1545	0.166	XP_001417613.1	147
3	XP_001419339.1	563	2	XP_001416295.1	1051	0.166	XP_001418105.1	78
3	XP_001419629.1	412	2	XP_001416688.1	260	0.166	XP_001418112.1	534
3	XP_001420239.1	3608	2	XP_001416790.1	163	0.166	XP_001418421.1	70
3	XP_001421421.1	401	2	XP_001416818.1	429	0.166	XP_001418862.1	145
3	XP_001421496.1	95	2	XP_001416870.1	1782	0.166	XP_001420232.1	68

Figure 3.17: Final output from the pipeline. Each column shows the score followed by protein ID and the total sequence length. The proteins present consistently across the three lists have higher likelihood to be present in the sample. There are two proteins (color-coded in red and violet) reported across the three lists, and one protein (color-coded in blue) reported in the first and third lists. The figure shows first 30 lines of each list.

From Figures 3.11, 3.14, 3.16, and 3.17 it can be observed that among the top protein hits, a number of proteins have identical scores. The report of such proteins was due to the m/z values of the associated peptides being within the range of the set threshold value and the pipeline finding the same number of peptides associated with the proteins.

For a given protein reported by the pipeline, the ratio of the number of identified peptides to the number of tryptic products possible can also be used to evaluate the pipeline. In particular, the ratio can indicate how much of the potential information is actually being used to identify peptides, and ultimately, plant proteins. For example, in Figure 3.18 the total count of peptides for protein XP_001417215.1 in the digested plant protein database is 224, and the total count of peptides which were used as evidence to report plant proteins is 13 (evidence peptides) with 10 PPM as the threshold value. The ratio of these two values is 0.0580, indicating that a small number potential peptides are being used to identify the proteins. A Perl script was written to collect the necessary information from the intermediate data files and calculate such ratios. Figure 3.18 shows protein names, peptide counts (evidence) and the total peptide count of selected input data. It also contains the aforementioned ratios.

Section 3.2 discussed how the charge state of the input non-plant peptides were predicted and used in the pipeline. Briefly, based on the counts of R, K and basic AAs only the highest possible charge was awarded to a peptide (up to some limits). However, in an earlier version of the pipeline multiple charges for each input peptides were allowed (2^+, 3^+, 4^+). When these peptides were used as input data, the number of false positive proteins increased. For example, Figure 3.19 reports the results on one test. In those results peptide TFGALSLTWK is reported twice, with charge states as 2^+, 3^+. Peptides that had charge as 3^+ were reported 3 times (peptide VFFCNSGTEANEGALKFARK is reported with charge states as 2^+, 3^+ and 4^+). Similarly peptides with 4^+ were reported 3 times (data not shown). This is the result of allowing multiple charges to the input peptides. The trial version of the pipeline favoured peptides with multiple charges, hence biasing the result towards those peptides having higher charge states. This resulted in report of erroneous proteins. Consequently, the single-most intense charge was used for all input data/peptides in the final version of the pipeline (Figure 3.20).

Protein_Name	Evi_cnt	Total_Cnt	Ratio
XP_001417215.1	13	224	0.0580
XP_001416048.1	6	615	0.0098
XP_001418911.1	6	377	0.0159
XP_001416378.1	5	2242	0.0022
XP_001418450.1	5	233	0.0215
XP_001421646.1	5	109	0.0459
XP_001417304.1	4	531	0.0075
XP_001420501.1	4	13	0.3077
XP_001421699.1	4	132	0.0303
XP_001422420.1	4	326	0.0123
YP_636192.1	4	446	0.0090
NP_084773.1	3	356	0.0084
XP_001415435.1	3	51	0.0588
XP_001416295.1	3	139	0.0216
XP_001416296.1	3	214	0.0140
XP_001416870.1	3	346	0.0087
XP_001417056.1	3	163	0.0184
XP_001417211.1	3	182	0.0165
XP_001417493.1	3	193	0.0155
XP_001417762.1	3	111	0.0270
XP_001417978.1	3	147	0.0204
XP_001418147.1	3	115	0.0261
XP_001418170.1	3	29	0.1034
XP_001418654.1	3	55	0.0545
XP_001418751.1	3	208	0.0144
XP_001419339.1	3	90	0.0333
XP_001419629.1	3	79	0.0380
XP_001420239.1	3	657	0.0046
XP_001421421.1	3	51	0.0588
XP_001421496.1	3	21	0.1429

Figure 3.18: Ratio of the number of identified peptides to the number of tryptic products. The values under Evi_cnt are the count of peptides which were used as evidence to report plant proteins. The values under Total_cnt are the total counts of peptides after *in silico* digestion with trypsin. The values under Ratio indicate the proportion of potential peptides used for protein identification and are calculated as the total count of peptides divided by the actual count of peptides used by the pipeline for protein identification.

(1)

SLHFFLAAWPVVGIWFTALGISTMAFNLNGFNFNQSVVDSQ
ANLGMEVMHER
SIFLFKK
TSLFSFR
MEHFGIMYPGFFRK
QSSSLPLLSSGTFLERIIFSRK
TIWFFMDPLMHYVR

(2)

881.5372	221.3915	4	3	SIFLFKK
856.4441	286.4886	3	2	TSLFSFR
881.5372	294.8529	3	3	SIFLFKK
1285.5902	322.4047	4	3	ANLGMEVMHER
1122.6069	375.2095	3	2	TFGALSLTWK
1563.8839	391.9782	4	5	SSLTEKLAGHPRLR
856.4441	429.2292	2	2	TSLFSFR

(3)

24	XP_001422626.1
23	YP_874717.1
15	XP_001416378.1
10	XP_001421255.1
9	XP_001417304.1

(4)

Evidence for protein XP_001422626.1:

125	134	1122.6069	375.2095	3	2	XP_001422626.1	TFGALSLTWK
302	315	1563.8839	391.9782	4	5	XP_001422626.1	SSLTEKLAGHPRLR
284	296	1417.7197	473.5804	3	2	XP_001422626.1	VSDPAFLANVTER
352	370	2032.2036	509.0581	4	3	XP_001422626.1	GDILRLVPPLIVSSAQVQK
106	124	2063.9079	516.9842	4	3	XP_001422626.1	NAGDAAWETVSFENGFHGR
75	94	2188.0727	548.0254	4	4	XP_001422626.1	VFFCNSGTEANEGALKFARK
125	134	1122.6069	562.3106	2	2	XP_001422626.1	TFGALSLTWK
352	370	2032.2036	678.4084	3	3	XP_001422626.1	GDILRLVPPLIVSSAQVQK
106	124	2063.9079	688.9765	3	3	XP_001422626.1	NAGDAAWETVSFENGFHGR
139	158	2091.0303	698.0173	3	2	XP_001422626.1	APFAPGLPGNTFTPYGDLEK
284	296	1417.7197	709.8670	2	2	XP_001422626.1	VSDPAFLANVTER
75	94	2188.0727	730.3648	3	4	XP_001422626.1	VFFCNSGTEANEGALKFARK
254	283	2931.3997	733.8571	4	3	XP_001422626.1	VAAVMAAGDHGSTFAGGPLVCAVANEVFDR
46	74	3172.6252	794.1635	4	6	XP_001422626.1	TLTHTSNLYHTEPGATLARKLTATSFADR
168	191	2567.2355	856.7524	3	2	XP_001422626.1	TCAVFVEPVQGEGGIYPADAEFLR
227	251	2567.3149	856.7788	3	3	XP_001422626.1	DAEPDMMSVAKPLANGLPIGAVLMK
254	283	2931.3997	978.1404	3	3	XP_001422626.1	VAAVMAAGDHGSTFAGGPLVCAVANEVFDR
352	370	2032.2036	1017.1090	2	3	XP_001422626.1	GDILRLVPPLIVSSAQVQK
106	124	2063.9079	1032.9611	2	3	XP_001422626.1	NAGDAAWETVSFENGFHGR
139	158	2091.0303	1046.5223	2	2	XP_001422626.1	APFAPGLPGNTFTPYGDLEK
75	94	2188.0727	1095.0435	2	4	XP_001422626.1	VFFCNSGTEANEGALKFARK
168	191	2567.2355	1284.6249	2	2	XP_001422626.1	TCAVFVEPVQGEGGIYPADAEFLR
227	251	2567.3149	1284.6646	2	3	XP_001422626.1	DAEPDMMSVAKPLANGLPIGAVLMK
254	283	2931.3997	1466.7070	2	3	XP_001422626.1	VAAVMAAGDHGSTFAGGPLVCAVANEVFD

Figure 3.19: Report of false positives when multiple charge states were allowed for each peptide. Plant peptides were used as input to the pipeline (1). With each peptide various values were associated (2) (see Figure 3.9 for explanation of the columns). A list of protein IDs with scores was the output (only a portion of the output is shown). This list was sorted based on the score associated with each protein (3) (the meaning of the columns is as in Figure 3.11). Upon analysis of the reported evidence (peptides), it was found that identical peptides with different m/z values were reported (4) (see Figure 3.12 for an explanation of the columns). As a result, the single most probable charge state was assigned to each peptide in the final pipeline.

(1)

11 XP_001422626.1
10 YP_874717.1
8 XP_001416378.1
4 XP_001417790.1
4 XP_001422461.1

(2)

Evidence for protein XP_001422626.1:

125	134	1122.6069	375.2095	3	2	XP_001422626.1	TFGALSLTW
302	315	1563.8839	391.9782	4	5	XP_001422626.1	SSLTEKLAGHPRLR
284	296	1417.7197	473.5804	3	2	XP_001422626.1	VSDPAFLANVTER
352	370	2032.2036	509.0581	4	3	XP_001422626.1	GDILRLVPPLIVSSAQVQK
106	124	2063.9079	516.9842	4	3	XP_001422626.1	NAGDAAWETVSFENGFHGR
75	94	2188.0727	548.0254	4	4	XP_001422626.1	VFFCNSGTEANEGALKFARK
139	158	2091.0303	698.0173	3	2	XP_001422626.1	APFAPGLPGNTFTPYGDLEK
254	283	2931.3997	733.8571	4	3	XP_001422626.1	VAAVMAAGDHGSTFAGGPLVCAVANEVFDR
46	74	3172.6252	794.1635	4	6	XP_001422626.1	TLTHTSNLYHTEPGATLARKLTATSFAD
168	191	2567.2355	856.7524	3	2	XP_001422626.1	TCAVFVEPVQGEGGIYPADAEFLR
227	251	2567.3149	856.7788	3	3	XP_001422626.1	DAEPDMMSVAKPLANGLPIGAVLMK

Figure 3.20: Report of proteins after compensating for false positives due to multiple charges on the peptides. The input to the pipeline remained the same as shown in Figure 3.19. However, in this case the charge state of each peptide was restricted. Only one charge state (the highest possible charge for each peptide) was allowed. This reduced the score associated with each protein (1). Upon analysis of the reported evidence (peptide), the peptides only carried the highest possible charge and the total count of peptides associated with that proteins was reduced (2), as reflected in the score associated with the proteins returned by the pipeline(1). Bottom of the figure (2), from left-to-right, the following information is provided in columns: Start, End, Molecular_Weight, m/z, Basic_AA, Charge, Sequence. For the meaning of these column headings refer to Figure 3.6.

3.5 Testing

The pipeline was tested and the outputs were analyzed. The test cases presented in this section show the preliminary results from the pipeline. The results presented here are solely from the testing of the pipeline. In the following sections, unless otherwise stated, "the list" refers to the initial potential plant protein list (for example, Figure 3.11 shows the initial list). Proteins which were to appear in the output lists (i.e. "known proteins") are color-coded in the figures.

There were three main purposes to the testing of the pipeline. The first purpose was to ensure that the behaviour of the pipeline is stable and that its output does not diverge; i.e. for a given set of known plant peptides or a combination of known plant and non-plant peptides along with associated m/z values as input, the pipeline should report similar plant peptides. The second purpose was to determine workable settings for the parameters of the pipeline: 1) the threshold value (e.g. 100, 50, 10 PPM), and 2) the number of peptides per input protein. The third purpose was to gather data to determine the error statistics (sensitivity, positive predictive value) for the pipeline. Usually sensitivity along with specificity are calculated as metrics. There is often a trade-off between sensitivity and specificity. Sensitivity and specificity are calculated using quantities such as false positives, false negatives, true positives, and true negatives. False positive (FP) refers to a plant protein reported by the pipeline that was not represented in the input data set. False negative (FN) refers to a plant protein that was represented in the input data set but was not reported by the pipeline. True positive refers to a plant protein represented in the input data set that was correctly identified and reported by the pipeline. Finally, a true negative (TN) is a non-plant protein in the input data set which was not reported by the pipeline. Protein identification programs typically try to minimize the reports of false positives (FPs) at the cost of more false negatives (FNs). However, sometimes it is desirable to have fewer false negative (FN) proteins reported at the risk of some false positives.

In order to determine specificity, calculated as TN/TN+FP, the count of true negatives (TNs) is necessary. In the experimental and test cases here, a true negative (TN) value is not meaningful. In particular, the pipeline can never output a non-plant protein. Therefore, any non-plant protein given as input will never be reported in the output. It is unrealistic to consider these non-results as a measure of the error rate of the program since, not being in the database, the non-plant proteins could never be reported.

Since specificity cannot be calculated, positive predictive value (PPV) is used instead. The PPV can be calculated as TP/(TP+FP) (Table 3.3). Sensitivity can be calculated as TP/(TP+FN) (Table 3.3). For the calculations of sensitivity and PPV only the top 30 hits were considered. For test cases where only plant protein peptides or a combination of plant and non-plant peptides were used as input, counts of TPs, FPs and FNs can be reported. For test cases where only non-plant

peptides were used as input, only FPs can be reported. No other values can be calculated as in this case the pipeline is forced to report false positives. Details of TPs, FPs and FNs for all test cases are shown using tables at the end of the section along with the summary reports that includes TPs, FPs, FNs, sensitivity and PPV. Tables showing calculations of sensitivity and PPV are given in the Appendix A. Before that, however, we describe the test cases.

Table 3.3: Formulae for calculating sensitivity and positive predictive value

Sensitivity	**Positive Predictive Value**
TP/(TP+FN)	TP/(TP+FP)

Test Cases

The data for error statistics was gathered from all the test cases and used in calculations of sensitivity and PPV.

In the following two test cases stability of the pipeline and workable parameters (different threshold values and number of peptides per proteins) were tested. 100, 50 and 10 PPM were the three threshold values used to test the reporting of plant proteins across the three lists. The same threshold values were used in test case 2, whereas the total number of peptides per protein, the protein counts, and the peptide count were changed.

Peptides from known plant proteins were used as input to the pipeline. The pipeline should report the known proteins among the top hits in the output lists. The plant proteins were selected in no specific order from the filtered plant database (Section 3.2.1.1).

Test 1: The three known plant proteins YP_874716.1, YP_874717.1, and XP_001422626.1 were selected from the database. A total of 24 peptides (selected in a non-specific manner) were used as input of which 2 were from YP_874716.1 and 11 were from each of YP_874717.1 and XP_001422626.1 (2+11+11). Figure A.3 in Appendix A shows a partial list of plant peptides.

When a threshold value of 100 PPM was used, proteins YP_874717.1 and XP_001422626.1 were reported among the top hits in at least two of the potential plant lists (Figure 3.21). These two proteins had more than 10 peptides each in the input set. The protein with only 2 peptides, YP_874716.1, was not among the top 30 hits reported. The plant protein YP_874717.1 (color-coded green) was reported consistently across the three lists. Protein XP_001422626.1 (red) was reported only in the first and second lists. The reported two proteins are true positive. Plant protein YP_874716.1 is a false negatives since it should have been reported and was not. Proteins XP_001416209.1, XP_001417105.1, XP_001417149.1, XP_001419214.1, XP_001420154.1, XP_001421360.1, NP_683826.1, NP_683843.1, NP_683862.1, NP_817146.1 and NP_862734.1 were also reported and were common to the first and the seconds lists. The latter group are all false positives. Tables 3.4 and 3.6 show the counts of TPs, FPs and FNs with 100 PPM as the threshold value.

11	XP_001422626.1	384	11	XP_001422626.1	384	0.3333	YP_717260.1	38
10	YP_874717.1	511	9	YP_874717.1	511	0.2500	YP_001001532.1	100
8	XP_001416378.1	18193	6	XP_001416378.1	18193	0.2500	YP_001152053.1	47
4	XP_001417790.1	1676	3	XP_001416177.1	4526	0.2500	YP_001152087.1	43
4	XP_001422461.1	1283	3	XP_001416209.1	717	0.2500	YP_001152210.1	50
3	XP_001416177.1	4526	3	XP_001416340.1	1213	0.2500	YP_874717.1	511
3	XP_001416209.1	717	3	XP_001417105.1	864	0.2222	XP_001418849.1	110
3	XP_001416340.1	1213	3	XP_001417149.1	667	0.2000	NP_817263.1	65
3	XP_001417105.1	864	3	XP_001417304.1	4003	0.2000	XP_001416392.1	56
3	XP_001417149.1	667	3	XP_001418766.1	1479	0.2000	XP_001421187.1	61
3	XP_001417304.1	4003	3	XP_001419214.1	552	0.2000	XP_001422208.1	56
3	XP_001418766.1	1479	3	XP_001420154.1	523	0.2000	XP_001422340.1	64
3	XP_001418796.1	1009	3	XP_001420401.1	2320	0.2000	YP_001152235.1	55
3	XP_001419214.1	552	3	XP_001420717.1	2267	0.1667	NP_084713.1	66
3	XP_001420154.1	523	3	XP_001420882.1	1292	0.1667	XP_001417541.1	67
3	XP_001420401.1	2320	3	XP_001421255.1	3191	0.1667	XP_001418912.1	75
3	XP_001420717.1	2267	3	XP_001421360.1	717	0.1667	XP_001420358.1	75
3	XP_001420882.1	1292	3	XP_001422461.1	1283	0.1667	XP_001421819.1	78
3	XP_001421255.1	3191	2	NP_683826.1	353	0.1667	YP_024398.1	66
3	XP_001421360.1	717	2	NP_683831.1	1627	0.1667	YP_717210.1	72
3	YP_024391.1	320	2	NP_683843.1	275	0.1667	YP_717225.1	70
3	YP_874665.1	320	2	NP_683861.1	1450	0.1429	NP_862819.1	89
2	NP_683826.1	353	2	NP_683862.1	458	0.1429	XP_001415389.1	91
2	NP_683831.1	1627	2	NP_817146.1	494	0.1429	XP_001416432.1	81
2	NP_683843.1	275	2	NP_862734.1	353	0.1429	XP_001416745.1	177
2	NP_683861.1	1450	2	NP_904209.1	353	0.1250	NP_862752.1	100
2	NP_683862.1	458	2	XP_001415360.1	546	0.1250	XP_001417193.1	96
2	NP_817146.1	494	2	XP_001415594.1	218	0.1250	XP_001417254.1	102
2	NP_862734.1	353	2	XP_001415630.1	530	0.1250	XP_001417413.1	100
2	NP_904209.1	353	2	XP_001415696.1	551	0.1250	XP_001417865.1	94

Figure 3.21: Result from the pipeline for known plant peptides used as input with 100 PPM as the threshold value — test case 1. The top 30 reported proteins are shown in the figure. From left-to-right: the first column lists the proteins based on m/z values (initial list), second column lists the proteins with unique peptides and the third column lists the proteins after correcting the bias of the pipeline for proteins with longer sequences. Known plant protein YP_874717.1 (in green) is reported in all three lists. Protein XP_001422626.1 (red) is reported in the first and second lists.

With 50 PPM as the threshold value known plant protein YP_874717.1 was common to the second and third lists (Figure 3.22). Protein XP_001416745.1 was present in all the three lists (a FP). Proteins XP_001420154.1, NP_683826.1, NP_862734.1, NP_904209.1, XP_001416209.1, XP_001416454.1, XP_001416667.1, XP_001416953.1, XP_001417105.1, XP_001417149.1, XP_001417366.1, XP_001417990.1 and XP_001418290.1 were also reported common in the first and second lists (all FPs). Two proteins XP_001415594.1 and XP_001415759.1 were common in the first and third lists (both FPs). Tables 3.4 and 3.6 show the counts of TPs, FPs and FNs with 50 PPM as the threshold value.

11	YP_717262.1	475	11	YP_636375.1	395	0.3333	YP_717260.1	38
10	YP_636405.1	344	9	YP_874717.1	511	0.2500	YP_001152087.1	43
4	XP_001416378.1	18193	3	XP_001416340.1	1213	0.2500	YP_001152210.1	50
4	XP_001417790.1	1676	3	XP_001416378.1	18193	0.2500	YP_874717.1	511
3	XP_001416340.1	1213	3	XP_001420154.1	523	0.2000	NP_817263.1	65
3	XP_001420154.1	523	3	XP_001420882.1	1292	0.2000	XP_001421187.1	61
3	XP_001420882.1	1292	3	XP_001421255.1	3191	0.2000	XP_001422340.1	64
3	XP_001421255.1	3191	2	NP_683826.1	353	0.1667	NP_084713.1	66
3	XP_001422461.1	1283	2	NP_683861.1	1450	0.1667	XP_001417541.1	67
2	NP_683826.1	353	2	NP_862734.1	353	0.1667	XP_001420358.1	75
2	NP_683861.1	1450	2	NP_904209.1	353	0.1667	XP_001421819.1	78
2	NP_862734.1	353	2	XP_001415594.1	218	0.1667	YP_024398.1	66
2	NP_904209.1	353	2	XP_001415759.1	219	0.1429	NP_862819.1	89
2	XP_001415594.1	218	2	XP_001416177.1	4526	0.1429	XP_001416432.1	81
2	XP_001415759.1	219	2	XP_001416209.1	717	0.1429	XP_001416745.1	177
2	XP_001416177.1	4526	2	XP_001416454.1	287	0.1250	XP_001417193.1	96
2	XP_001416209.1	717	2	XP_001416502.1	1189	0.1250	XP_001417254.1	102
2	XP_001416454.1	287	2	XP_001416563.1	1843	0.1250	XP_001417413.1	100
2	XP_001416502.1	1189	2	XP_001416667.1	656	0.1250	XP_001420217.1	95
2	XP_001416563.1	1843	2	XP_001416745.1	177	0.1250	YP_001001532.1	100
2	XP_001416667.1	656	2	XP_001416953.1	459	0.1250	YP_001001596.1	93
2	XP_001416745.1	177	2	XP_001417017.1	1393	0.1250	YP_001294313.1	92
2	XP_001416953.1	459	2	XP_001417046.1	1252	0.1250	YP_001294416.1	93
2	XP_001417017.1	1393	2	XP_001417105.1	864	0.1250	YP_024358.1	93
2	XP_001417046.1	1252	2	XP_001417149.1	667	0.1250	YP_358646.1	93
2	XP_001417105.1	864	2	XP_001417366.1	415	0.1250	YP_778554.1	93
2	XP_001417149.1	667	2	XP_001417990.1	211	0.1250	YP_874712.1	93
2	XP_001417366.1	415	2	XP_001418290.1	515	0.1250	YP_874745.1	94
2	XP_001417990.1	211	2	XP_001418414.1	864	0.1176	XP_001415594.1	218
2	XP_001418290.1	515	2	XP_001418754.1	2823	0.1176	XP_001415759.1	219

Figure 3.22: Result from the pipeline for known plant peptides used as input with 50 PPM as the threshold value — test case 1. Top 30 reported proteins are shown in the figure. Annotation for each column is same as Figure 3.21. Known plant protein YP_874717.1 (color-coded in blue) is reported in the second and third lists.

With 10 PPM as the threshold value known protein YP_874717.1 (color-coded in green) was reported in the all the three lists and another known plant protein XP_001422626.1 (orange) was reported in the first and second lists (Figure 3.23). Proteins NP_683826.1, NP_862734.1, NP_904209.1,

YP_001001514.1, YP_001004166.1, YP_001123011.1, YP_001123531.1, YP_001294250.1, YP_024363.1, YP_358554.1, YP_538828.1, YP_636279.1, YP_762241.1, YP_784453.1 and YP_874633.1 were also reported and were common to all the three lists (all FPs). Proteins XP_001415777.1, XP_001415685.1, XP_001415532.1, XP_001415530.1, XP_001415509.1 and NP_084672.1 were common to the first and second lists (all FPs). Tables 3.4 and 3.6 show the counts of TPs, FPs and FNs with 10 PPM as the threshold value.

10	XP_001422626.1	384	10	XP_001422626.1	384	0.2500	**YP_874717.1**	511
10	**YP_874717.1**	511	9	**YP_874717.1**	511	0.1111	XP_001417526.1	114
2	NP_683826.1	353	2	NP_683826.1	353	0.1111	XP_001419213.1	108
2	NP_862734.1	353	2	NP_862734.1	353	0.1111	XP_001419514.1	115
2	NP_904209.1	353	2	NP_904209.1	353	0.1000	XP_001418338.1	130
2	XP_001418796.1	1009	2	XP_001421255.1	3191	0.1000	XP_001421058.1	125
2	XP_001421255.1	3191	2	YP_001001514.1	353	0.1000	XP_001421077.1	225
2	YP_001001514.1	353	2	YP_001004166.1	353	0.0909	XP_001415892.1	141
2	YP_001004166.1	353	2	YP_001123011.1	353	0.0909	XP_001416495.1	132
2	YP_001123011.1	353	2	YP_001123531.1	353	0.0769	XP_001422409.1	166
2	YP_001123531.1	353	2	YP_001294250.1	353	0.0714	NP_683826.1	353
2	YP_001294250.1	353	2	YP_024363.1	353	0.0714	NP_862734.1	353
2	YP_024363.1	353	2	YP_358554.1	353	0.0714	NP_904209.1	353
2	YP_358554.1	353	2	YP_538828.1	353	0.0714	XP_001416745.1	177
2	YP_538828.1	353	2	YP_636279.1	353	0.0714	XP_001418633.1	177
2	YP_636279.1	353	2	YP_762241.1	353	0.0714	XP_001421814.1	175
2	YP_762241.1	353	2	YP_784453.1	353	0.0714	XP_001422504.1	177
2	YP_784453.1	353	2	YP_874633.1	353	0.0714	YP_001001514.1	353
2	YP_874633.1	353	2	XP_001418530.1	1034	0.0714	YP_001004166.1	353
2	XP_001418530.1	1034	1	NP_084672.1	751	0.0714	YP_001123011.1	353
1	NP_084672.1	751	1	NP_683831.1	1627	0.0714	YP_001123531.1	353
1	NP_683831.1	1627	1	NP_817132.1	353	0.0714	YP_001294250.1	353
1	NP_817132.1	353	1	NP_862754.1	750	0.0714	YP_024363.1	353
1	NP_862754.1	750	1	XP_001415509.1	753	0.0714	YP_358554.1	353
1	XP_001415509.1	753	1	XP_001415530.1	355	0.0714	YP_538828.1	353
1	XP_001415530.1	355	1	XP_001415532.1	688	0.0714	YP_636279.1	353
1	XP_001415532.1	688	1	XP_001415599.1	1503	0.0714	YP_762241.1	353
1	XP_001415599.1	1503	1	XP_001415685.1	288	0.0714	YP_784453.1	353
1	XP_001415685.1	288	1	XP_001415777.1	935	0.0714	YP_874633.1	353
1	XP_001415777.1	935	1	XP_001415811.1	484	0.0714	XP_001418530.1	1034

Figure 3.23: Result from the pipeline for known plant peptides used as input with 10 PPM as the threshold value — test case 1. Top 30 reported proteins are shown in the figure. Annotation for each column is same as in Figure 3.21. Known plant protein YP_874717.1 (color-coded in green) is reported in the all the lists. Plant protein XP_001422626.1 (color-coded in orange) is reported in all the first and second lists.

Test 2: Peptides from the known plants were used as input with a threshold value of 100 PPM. A total of 16 peptides were used as input, 4 peptides each from the known plant proteins (4+4+4+4). The peptides were selected in a non-specific manner from the following proteins: YP_636294.1, YP_636291.1, YP_717254.1 and YP_717251.1. In this case, the total count of peptides was less than the number of peptides used in test case 1.

Only one protein (YP_717251.1) was reported consistently in the three lists (Figure 3.24). All the 4 input proteins were reported across at least 2 lists. Proteins XP_001419520.1, XP_001419587.1, YP_762256.1, NP_084679.1, NP_862749.1, XP_001415413.1, XP_001415642.1 and XP_001415675.1 were common to the first and the seconds lists (all FPs). Tables 3.4 and 3.6 show the counts of TPs, FPs and FNs with 100 PPM as the threshold value.

13	XP_001416378.1	18193	8	XP_001416378.1	18193	0.5000	**YP_717251.1**	100
13	XP_001417215.1	1345	6	XP_001416023.1	1918	0.4000	XP_001415824.1	63
11	XP_001420239.1	3608	6	XP_001418571.1	4390	0.3333	XP_001416152.1	35
6	XP_001416023.1	1918	5	XP_001418754.1	2823	0.2857	YP_001152218.1	82
6	XP_001418571.1	4390	5	XP_001420618.1	1899	0.2857	YP_358636.1	91
5	XP_001418754.1	2823	5	XP_001421248.1	4434	0.2500	NP_817155.1	44
5	XP_001420618.1	1899	5	XP_001422532.1	3060	0.2500	XP_001415423.1	51
5	XP_001421248.1	4434	5	YP_636177.1	3462	0.2500	XP_001418346.1	95
5	XP_001422532.1	3060	5	YP_636291.1	1072	0.2500	XP_001419210.1	49
5	YP_636177.1	3462	4	XP_001415523.1	587	0.2500	YP_001123599.1	52
5	YP_636291.1	1072	4	XP_001416048.1	4962	0.2500	YP_001152134.1	41
4	XP_001415523.1	587	4	XP_001417143.1	1170	0.2500	YP_762308.1	52
4	XP_001416048.1	4962	4	XP_001419520.1	470	0.2222	XP_001420451.1	111
4	XP_001417143.1	1170	4	XP_001419587.1	824	0.2222	YP_636187.1	109
4	XP_001418911.1	1823	4	XP_001420022.1	1148	0.2000	NP_862803.1	57
4	XP_001419520.1	470	4	XP_001420487.1	2198	0.2000	XP_001419951.1	57
4	XP_001419587.1	824	4	XP_001421896.1	1448	0.2000	XP_001420920.1	65
4	XP_001420022.1	1148	4	XP_001422084.1	1010	0.2000	XP_001422327.1	59
4	XP_001420487.1	2198	4	YP_001294179.1	353	0.2000	XP_001422493.1	59
4	XP_001421896.1	1448	4	YP_538843.1	353	0.2000	YP_001152142.1	55
4	XP_001422084.1	1010	4	YP_636294.1	353	0.2000	YP_001152196.1	62
4	YP_001294179.1	353	4	**YP_717251.1**	100	0.2000	YP_001152267.1	65
4	YP_538843.1	353	4	YP_717254.1	751	0.2000	YP_001294410.1	57
4	YP_636294.1	353	4	YP_762256.1	353	0.2000	YP_538871.1	128
4	**YP_717251.1**	100	3	NP_084679.1	353	0.2000	YP_538898.1	55
4	YP_717254.1	751	3	NP_084773.1	2280	0.2000	YP_784435.1	57
4	YP_762256.1	353	3	NP_862749.1	353	0.2000	YP_784530.1	54
3	NP_084679.1	353	3	XP_001415413.1	496	0.1875	XP_001418825.1	198
3	NP_084773.1	2280	3	XP_001415537.1	1062	0.1667	NP_817223.1	75
3	NP_862749.1	353	3	XP_001415642.1	424	0.1667	XP_001417050.1	78
3	XP_001415413.1	496	3	XP_001415675.1	356	0.1667	XP_001417458.1	73
3	XP_001415537.1	1062	3	XP_001416055.1	407	0.1667	XP_001418791.1	227
3	XP_001415642.1	424	3	XP_001416284.1	1545	0.1667	XP_001418912.1	75
3	XP_001415675.1	356	3	XP_001416350.1	740	0.1667	XP_001420354.1	73

Figure 3.24: Result from the pipeline for known plant peptides used as input with 100 PPM as the threshold value — test case 2. The top 30 reported proteins are shown in the figure. Annotation for each column is the same as in Figure 3.21. Protein YP_717251.1 (color-coded in sky blue) was observed consistently across the three potential plant list. Proteins YP_636294.1 (light-red), YP_636291.1 (red) and YP_717254.1 (blue) were reported in the first and second lists.

With 50 PPM as the threshold value known plant protein YP_717251.1 was common to the second and third lists (Figure 3.25) while protein YP_636294.1 was common to the first and second lists (Figure 3.25). Proteins XP_001419520.1, YP_001294179.1, YP_538843.1, YP_717245.1, YP_762256.1, NP_084679.1, NP_862749.1, XP_001417105.1, XP_001417151.1 and XP_001418825.1 were common to the first and seconds lists (all FPs). Tables 3.4 and 3.6 show the counts of TPs, FPs and FNs with 50 PPM as the threshold value.

13	XP_001417215.1	1345	5	XP_001416378.1	18193	0.5000	YP_717251.1	100
7	XP_001416378.1	18193	5	XP_001418571.1	4390	0.3333	XP_001416152.1	35
5	XP_001418571.1	4390	5	YP_636177.1	3462	0.2857	YP_001152218.1	82
5	YP_636177.1	3462	5	YP_636299.1	1073	0.2857	YP_358636.1	91
5	YP_636299.1	1073	4	XP_001415523.1	587	0.2500	NP_817155.1	44
4	XP_001415523.1	587	4	XP_001416048.1	4962	0.2500	XP_001415423.1	51
4	XP_001416048.1	4962	4	XP_001418754.1	2823	0.2500	YP_001123599.1	52
4	XP_001418754.1	2823	4	XP_001419520.1	470	0.2500	YP_762308.1	52
4	XP_001418911.1	1823	4	XP_001421248.1	4434	0.2000	NP_862803.1	57
4	XP_001419520.1	470	4	XP_001422084.1	1010	0.2000	XP_001415824.1	63
4	XP_001421248.1	4434	4	XP_001422532.1	3060	0.2000	XP_001419951.1	57
4	XP_001422084.1	1010	4	YP_001294179.1	353	0.2000	XP_001420920.1	65
4	XP_001422532.1	3060	4	YP_538843.1	353	0.2000	XP_001420924.1	363
4	YP_001294179.1	353	4	YP_636294.1	353	0.2000	YP_001152142.1	55
4	YP_538843.1	353	4	YP_717251.1	100	0.2000	YP_001152196.1	62
4	YP_636294.1	353	4	YP_717245.1	757	0.2000	YP_001294410.1	57
4	YP_712751.1	200	4	YP_762256.1	353	0.2000	YP_538871.1	128
4	YP_717245.1	757	3	NP_084679.1	353	0.2000	YP_538898.1	55
4	YP_762256.1	353	3	NP_862749.1	353	0.2000	YP_784435.1	57
3	NP_084679.1	353	3	XP_001416023.1	1918	0.2000	YP_784530.1	54
3	NP_862749.1	353	3	XP_001416284.1	1545	0.1875	XP_001418825.1	198
3	XP_001416023.1	1918	3	XP_001417092.1	3039	0.1667	XP_001417050.1	78
3	XP_001416284.1	1545	3	XP_001417105.1	864	0.1667	XP_001417458.1	73
3	XP_001417092.1	3039	3	XP_001417143.1	1170	0.1667	XP_001418912.1	75
3	XP_001417105.1	864	3	XP_001417151.1	565	0.1667	XP_001420354.1	73
3	XP_001417143.1	1170	3	XP_001418825.1	198	0.1667	YP_001001569.1	77
3	XP_001417151.1	565	3	XP_001418928.1	1213	0.1667	YP_001294219.1	77
3	XP_001418825.1	198	3	XP_001419372.1	1105	0.1667	YP_001294387.1	77
3	XP_001418928.1	1213	3	XP_001419587.1	824	0.1667	YP_717221.1	78
3	XP_001419372.1	1105	3	XP_001420008.1	1069	0.1667	YP_784508.1	77

Figure 3.25: Result from the pipeline for known plant peptides used as input with 50 PPM as the threshold value — test case 2. The top 30 reported proteins are shown in the figure. Annotation for each column is the same as in Figure 3.21. Protein YP_717251.1 (color-coded in red) was common to the second and third potential plant lists. Proteins YP_636294.1 (orange) was reported in the first and second lists.

With 10 PPM as the threshold value two known proteins YP_636294.1 (color-coded in orange) and YP_717251.1 (Figure 3.26) were reported in all the all the lists. The other two known plant proteins YP_636291.1 and YP_717254.1 were reported in the first and second lists (Figure 3.26). Proteins YP_001294179.1, YP_538843.1, NP_862749.1, YP_636260.1, YP_874638.1, NP_683829.1, NP_817259.1, NP_904202.1 and XP_001416511.1 were common to all the three lists (all FPs). Proteins YP_762256.1, NP_084679.1, YP_001004181.1, YP_001122941.1, YP_001123025.1, YP_001123194.1, YP_001123545.1, YP_001123720.1, YP_001294265.1, YP_001294347.1, YP_024367.1, YP_358573.1, YP_778485.1 and YP_784381.1 were common to the first and second lists (all FPs) . Tables 3.4 and 3.6 show the counts of TPs, FPs and FNs with 10 PPM as the threshold value.

4	YP_001294179.1	353	4	YP_001294179.1	353	0.5000	YP_717251.1	100
4	YP_538843.1	353	4	YP_538843.1	353	0.1667	XP_001417050.1	78
4	YP_636291.1	1072	4	YP_636291.1	1072	0.1667	XP_001420354.1	73
4	YP_636294.1	353	4	YP_636294.1	353	0.1429	NP_683860.1	89
4	YP_717251.1	100	4	YP_717251.1	100	0.1429	XP_001418671.1	88
4	YP_717254.1	751	4	YP_717254.1	751	0.1429	YP_001294179.1	353
4	YP_762256.1	353	4	YP_762256.1	353	0.1429	YP_538843.1	353
3	NP_084679.1	353	3	NP_084679.1	353	0.1429	YP_636294.1	353
3	NP_862749.1	353	3	NP_862749.1	353	0.1429	YP_762256.1	353
3	XP_001418754.1	2823	3	XP_001418754.1	2823	0.1250	XP_001417742.1	99
3	YP_001004181.1	353	3	YP_001004181.1	353	0.1250	XP_001422172.1	99
3	YP_001122941.1	353	3	YP_001122941.1	353	0.1111	XP_001416436.1	109
3	YP_001123025.1	353	3	YP_001123025.1	353	0.1111	XP_001416505.1	114
3	YP_001123194.1	353	3	YP_001123194.1	353	0.1111	XP_001421466.1	108
3	YP_001123545.1	353	3	YP_001123545.1	353	0.1111	YP_001152130.1	107
3	YP_001123720.1	353	3	YP_001123720.1	353	0.1111	YP_636237.1	105
3	YP_001294265.1	353	3	YP_001294265.1	353	0.1071	NP_084679.1	353
3	YP_001294347.1	353	3	YP_001294347.1	353	0.1071	NP_862749.1	353
3	YP_024367.1	353	3	YP_024367.1	353	0.1071	YP_001004181.1	353
3	YP_358573.1	353	3	YP_358573.1	353	0.1071	YP_001122941.1	353
3	YP_636177.1	3462	3	YP_636177.1	3462	0.1071	YP_001123025.1	353
3	YP_636260.1	751	3	YP_636260.1	751	0.1071	YP_001123194.1	353
3	YP_778485.1	353	3	YP_778485.1	353	0.1071	YP_001123545.1	353
3	YP_784381.1	353	3	YP_784381.1	353	0.1071	YP_001123720.1	353
3	YP_874638.1	353	3	YP_874638.1	353	0.1071	YP_001294265.1	353
2	NP_683829.1	748	2	NP_683829.1	748	0.1071	YP_001294347.1	353
2	NP_817259.1	353	2	NP_817259.1	353	0.1071	YP_024367.1	353
2	NP_904202.1	750	2	NP_904202.1	750	0.1071	YP_358573.1	353
2	XP_001416284.1	1545	2	XP_001416284.1	1545	0.1071	YP_778485.1	353
2	XP_001416511.1	322	2	XP_001416511.1	322	0.1071	YP_784381.1	353

Figure 3.26: Result from the pipeline for known plant peptides used as input with 10 PPM as the threshold value — test case 2. The top 30 reported proteins are shown in the figure. Annotation for each column is the same as in Figure 3.21. Proteins YP_717251.1 (color-coded in red) and YP_636294.1 (orange) were observed consistently across the three potential plant list. Proteins YP_636291.1 (blue) and YP_717254.1 (green) were reported in the first and second lists.

Repeating such tests (test case 1 and 2) three more times with varying numbers of known plant proteins (3-6) and varying numbers of peptides (8-10) from each protein yielded the following results (data not shown): all the known plant proteins were reported (in at least one of the three lists) among the first 30 hits when a threshold value of 100 PPM or 10 PPM was selected. If the threshold value was relaxed (for example, 50 instead of 10 PPM) more FPs were observed; i.e., the counts of TPs were reduced when compared to the TPs reported with the threshold value of 10 PPM. Typically at least 4 peptides from each plant protein were needed for the protein to be reported in the top 30 hits. If 2 to 3 peptides per protein were provided, proteins were reported by the pipeline, though not among the top 30 hits.

Test 3: In this case stability of the pipeline along with the input parameter (threshold value) were tested by providing mixed data set—known non-plant and known plant peptides—as the input and using 100, 50, and 10 PPM as the threshold value. The total count of peptides for each selected proteins was kept constant. No specific criterion was used for choosing proteins for the data set. A total of 24 peptides were used as input. The non-plant peptides (all from Pig) were selected from the following proteins (total count of non-plant peptides is shown within brackets following the Protein ID): 2ABB_PIG (4), A1AT_PIG (4) and ABA54553.1 (4). The plant peptides were from following proteins: YP_874717.1 (4), YP_874716.1 (4) and YP_874727.1 (4).

With 100 PPM as threshold value all the three plant proteins were reported. The known plant proteins are differentially color-coded in Figure 3.27. Known plant proteins YP_874717.1, YP_874716.1 and YP_874727.1 were common to the first and second lists (Figure 3.27). Protein YP_778511.1 was common to the first and third lists (a FP). Proteins XP_001415516.1, XP_001415551.1, XP_001416241.1, XP_001416626.1, XP_001417857.1, XP_001417965.1, XP_001420166.1, XP_001420614.1, XP_001421252.1, XP_001422003.1, XP_001422780.1, YP_024363.1, YP_024372.1, YP_874644.1 and NP_683826.1 were also reported and were common in the first and second lists (all FPs). Tables 3.4 and 3.6 show the counts of TPs, FPs and FNs with 100 PPM as the threshold value.

14	XP_001416378.1	18193	8	XP_001416378.1	18193	0.4000	XP_001417070.1	60
6	YP_778511.1	231	5	XP_001415516.1	724	0.3333	NP_084682.1	34
5	XP_001415516.1	724	4	XP_001417293.1	2759	0.3333	NP_862748.1	34
4	XP_001417293.1	2759	4	YP_874717.1	511	0.3333	YP_001004180.1	34
4	XP_001417790.1	1676	3	XP_001415551.1	742	0.3333	YP_001123719.1	34
4	YP_874717.1	511	3	XP_001416241.1	995	0.3333	YP_024369.1	34
3	XP_001415551.1	742	3	XP_001416626.1	714	0.3333	YP_717208.1	37
3	XP_001416241.1	995	3	XP_001417857.1	794	0.3333	YP_778511.1	231
3	XP_001416626.1	714	3	XP_001417965.1	609	0.2500	XP_001415885.1	49
3	XP_001417857.1	794	3	XP_001418362.1	1432	0.2500	XP_001416183.1	52
3	XP_001417965.1	609	3	XP_001418571.1	4390	0.2500	XP_001416730.1	51
3	XP_001418362.1	1432	3	XP_001418754.1	2823	0.2500	YP_001152058.1	47
3	XP_001418571.1	4390	3	XP_001420166.1	339	0.2500	YP_024357.1	46
3	XP_001418754.1	2823	3	XP_001420614.1	880	0.2500	YP_636246.1	42
3	XP_001420166.1	339	3	XP_001421248.1	4434	0.2500	YP_762308.1	52
3	XP_001420614.1	880	3	XP_001421252.1	292	0.2000	XP_001421735.1	59
3	XP_001421248.1	4434	3	XP_001422003.1	481	0.2000	YP_024336.1	59
3	XP_001421252.1	292	3	XP_001422780.1	936	0.2000	YP_784435.1	57
3	XP_001422003.1	481	3	YP_001001592.1	1868	0.2000	YP_874700.1	59
3	XP_001422780.1	936	3	YP_001152259.1	2062	0.2000	YP_874785.1	61
3	YP_001001592.1	1868	3	YP_024363.1	353	0.1818	YP_024409.1	143
3	YP_001152259.1	2062	3	YP_024372.1	683	0.1818	YP_874685.1	143
3	YP_024363.1	353	3	YP_874644.1	682	0.1667	NP_817243.1	73
3	YP_024372.1	683	3	YP_874716.1	353	0.1667	XP_001417379.1	75
3	YP_874644.1	682	3	YP_874727.1	682	0.1667	XP_001418086.1	77
3	YP_874716.1	353	2	NP_084773.1	2280	0.1667	XP_001422542.1	146
3	YP_874727.1	682	2	NP_683776.1	1373	0.1667	YP_001294219.1	77
2	NP_084773.1	2280	2	NP_683826.1	353	0.1667	YP_717267.1	74
2	NP_683776.1	1373	2	NP_817153.1	1209	0.1538	XP_001416525.1	168
2	NP_683826.1	353	2	NP_862734.1	353	0.1429	NP_689364.1	91

Figure 3.27: Result from the pipeline for a mixture of known plant and non-plant peptides used as input with 100 PPM as the threshold value — test 3. The top 30 reported proteins are shown in the figure. Annotation for each column is the same as in Figure 3.21. The three known plant proteins, YP_874717.1 (color-coded in green), YP_874716.1 (cyan) and YP_874727.1 (red) were reported in the first two lists.

If 50 PPM was selected as the threshold value, known plant protein YP_874716.1 was reported in the second list (Figure 3.28). Proteins NP_904209.1, XP_001415516.1, XP_001415551.1, XP_001416626.1, XP_001417965.1, XP_001420614.1, XP_001421252.1, YP_024363.1, YP_024372.1, YP_874644.1, XP_001417535.1, XP_001417284.1, NP_683826.1, NP_862734.1, XP_001415862.1, XP_001415888.1, XP_001416126.1, XP_001415959.1, XP_001416209.1, XP_001416241.1 and XP_001416267.1 were also reported and were common to the first and second lists (all FPs) . Proteins YP_778511.1 and XP_001416525.1 were common in the first and third lists (all FPs). Tables 3.4 and 3.6 show the counts of TPs, FPs and FNs with 50 PPM as the threshold value.

12	XP_001416378.1	18193	6	XP_001416378.1	18193	0.4000	XP_001417070.1	60
6	YP_778511.1	231	4	YP_874716.1	353	0.3333	YP_717208.1	37
4	XP_001417790.1	1676	3	XP_001415516.1	724	0.3333	YP_778511.1	231
4	NP_904209.1	353	3	XP_001415551.1	742	0.2500	XP_001415885.1	49
3	XP_001415516.1	724	3	XP_001416626.1	714	0.2500	XP_001416183.1	52
3	XP_001415551.1	742	3	XP_001417965.1	609	0.2500	YP_024357.1	46
3	XP_001416626.1	714	3	XP_001418571.1	4390	0.2500	YP_636246.1	42
3	XP_001417965.1	609	3	XP_001420614.1	880	0.2500	YP_762308.1	52
3	XP_001418571.1	4390	3	XP_001421252.1	292	0.2000	XP_001421735.1	59
3	XP_001420614.1	880	3	YP_024363.1	353	0.2000	YP_784435.1	57
3	XP_001421252.1	292	3	YP_024372.1	683	0.1818	YP_024409.1	143
3	YP_024363.1	353	3	YP_874644.1	682	0.1818	YP_874685.1	143
3	YP_024372.1	683	3	XP_001416907.1	654	0.1667	XP_001418086.1	77
3	YP_874644.1	682	3	XP_001418286.1	399	0.1667	XP_001422542.1	146
3	XP_001417535.1	244	2	NP_683826.1	353	0.1667	YP_001294219.1	77
3	XP_001417284.1	432	2	NP_862734.1	353	0.1667	YP_717267.1	74
2	NP_683826.1	353	2	NP_904209.1	353	0.1538	XP_001416525.1	168
2	NP_862734.1	353	2	XP_001415739.1	3600	0.1429	NP_689364.1	91
2	NP_904200.1	353	2	XP_001415862.1	983	0.1429	XP_001416463.1	91
2	XP_001415739.1	3600	2	XP_001415888.1	315	0.1429	XP_001417204.1	79
2	XP_001415862.1	983	2	XP_001415959.1	254	0.1429	XP_001418041.1	178
2	XP_001415888.1	315	2	XP_001416012.1	1307	0.1429	XP_001422438.1	88
2	XP_001415959.1	254	2	XP_001416126.1	378	0.1429	YP_001294305.1	80
2	XP_001416012.1	1307	2	XP_001416209.1	717	0.1304	XP_001421252.1	292
2	XP_001416126.1	378	2	XP_001416241.1	995	0.1250	XP_001416807.1	101
2	XP_001416209.1	717	2	XP_001416267.1	364	0.1250	XP_001417413.1	100
2	XP_001416241.1	995	2	XP_001416502.1	1189	0.1250	XP_001421189.1	98
2	XP_001416267.1	364	2	XP_001416525.1	168	0.1250	YP_636461.1	102
2	XP_001416502.1	1189	2	XP_001416563.1	1843	0.1250	YP_717237.1	92
2	XP_001416525.1	168	2	XP_001416907.1	654	0.1111	NP_689355.1	116

Figure 3.28: Result from the pipeline for a mixture of known plant and non-plant peptides used as input with 50 PPM as the threshold value — test 3. The top 30 reported proteins are shown in the figure. Annotation for each column is the same as in Figure 3.21. Known plant protein YP_874716.1 (color-coded in magenta) was reported in the second list.

With 10 PPM known plant proteins YP_874717.1 and YP_874716.1 were reported in all the three lists (Figure 3.29). Plant protein YP_874727.1 was reported in the first and second lists (Figure 3.29). Proteins YP_024363.1, YP_024372.1, XP_001415576.1, XP_001416265.1, XP_001416126.1, YP_636279.1, YP_762241.1, YP_784453.1, YP_358554.1, YP_874633.1 and YP_538828.1 were also reported and were common in the first and second lists (all FPs). Proteins YP_024409.1, NP_689377.1 and NP_689355.1 were common to the first and third lists (all FPs). Finally, proteins NP_683826.1, NP_862734.1, NP_904209.1, XP_001421158.1, XP_001416525.1, XP_001418041.1 and NP_689355.1 were common to all the three lists (all FPs). Tables 3.4 and 3.6 show the counts of TPs, FPs and FNs with 10 PPM as the threshold value.

4	YP_874717.1	511	4	YP_874717.1	511	0.2500	XP_001416183.1	52
3	YP_024363.1	353	3	YP_024363.1	353	0.2500	YP_024357.1	46
3	YP_024372.1	683	3	YP_024372.1	683	0.2500	YP_762308.1	52
3	YP_874644.1	682	3	YP_874644.1	682	0.2000	YP_784435.1	57
3	YP_874716.1	353	3	YP_874716.1	353	0.1818	YP_024409.1	143
3	YP_874727.1	682	3	YP_874727.1	682	0.1818	YP_874685.1	143
2	NP_683826.1	353	2	NP_683826.1	353	0.1538	XP_001416525.1	168
2	NP_862734.1	353	2	NP_862734.1	353	0.1429	XP_001417204.1	79
2	NP_904209.1	353	2	NP_904209.1	353	0.1429	XP_001418041.1	178
2	XP_001416126.1	378	2	XP_001416126.1	378	0.1111	NP_689355.1	116
2	XP_001416378.1	18193	2	XP_001416378.1	18193	0.1111	XP_001419388.1	110
2	XP_001416525.1	168	2	XP_001416525.1	168	0.1071	YP_024363.1	353
2	XP_001418041.1	178	2	XP_001418041.1	178	0.1071	YP_874716.1	353
2	XP_001421158.1	281	2	XP_001421158.1	281	0.1000	NP_689377.1	126
2	XP_001422675.1	638	2	XP_001422675.1	638	0.1000	XP_001420603.1	130
2	YP_001001514.1	353	2	YP_001001514.1	353	0.1000	YP_001152214.1	124
2	YP_001004166.1	353	2	YP_001004166.1	353	0.1000	YP_588280.1	126
2	YP_001123011.1	353	2	YP_001123011.1	353	0.1000	YP_874717.1	511
2	YP_001123531.1	353	2	YP_001123531.1	353	0.0909	XP_001416306.1	136
2	YP_001294250.1	353	2	YP_001294250.1	353	0.0909	XP_001416883.1	137
2	YP_024409.1	143	2	YP_024409.1	143	0.0909	XP_001421158.1	281
2	YP_358554.1	353	2	YP_358554.1	353	0.0833	XP_001421230.1	150
2	YP_538828.1	353	2	YP_538828.1	353	0.0833	XP_001421906.1	146
2	YP_636279.1	353	2	YP_636279.1	353	0.0833	XP_001422542.1	146
2	YP_762241.1	353	2	YP_762241.1	353	0.0769	XP_001422409.1	166
2	YP_784453.1	353	2	YP_784453.1	353	0.0714	NP_683826.1	353
2	YP_874633.1	353	2	YP_874633.1	353	0.0714	NP_862734.1	353
2	YP_874685.1	143	2	YP_874685.1	143	0.0714	NP_904209.1	353
1	NP_689355.1	116	1	NP_689355.1	116	0.0714	XP_001419935.1	172
1	NP_689377.1	126	1	NP_689377.1	126	0.0714	XP_001422717.1	180

Figure 3.29: Result from the pipeline for a mixture of known plant and non-plant peptides used as input with 10 PPM as the threshold value — test 3. The top 30 reported proteins are shown in the figure. Annotation for each column is the same as in Figure 3.21. Known plant proteins, YP_874717.1 and YP_874716.1 (color-coded in orange, sky-blue respectively) were reported in all the three lists. Known plant protein YP_874727.1 (navy blue) was reported in the first and second lists.

In the next two cases, known non-plant peptides were used as input to the pipeline. In these cases, the aim was to estimate the background level of false positives for the pipeline. Note that in this case the pipeline is forced to report proteins from a plant-only database. All the reported proteins will be false positives. The TP count should be at minimum.

Test 4: A total of 20 peptides from 4 non-plant proteins (Pig) were selected as input with 5 peptides from each protein. Proteins and peptides were selected in no specific order. The proteins were 2ABB_PIG, 5HT1B_PIG, A1AT_PIG and AAKG1_PIG. A threshold value of 100 PPM was used.

Protein NP_683769.1 appeared consistently across the three potential plant lists (Figure 3.30). Protein YP_778511.1 was common to the first and the third lists. Furthermore, proteins XP_001417211.1, XP_001421657.1, XP_001416600.1, and XP_001420157.1 were common to the first and second lists (all FPs). Tables 3.4 and 3.6 show the counts of TPs, FPs and FNs with 100 PPM as the threshold value.

Test 5: The experiment in test 4 was repeated with the modification that the non-plant species was mouse, the count of proteins was reduced, and the count of peptides was both increased and decreased. A total of 23 peptides from 3 non-plant proteins from mouse were used as input and a threshold value of 100 PPM was selected. The peptides and proteins were selected in no specific order. Nine peptides from 2A5A_MOUSE, four from 1433T_MOUSE and ten from A2M_MOUSE were used as input.

In the results, protein XP_001420313.1 was common to all the three lists (Figure 3.31). Fourteen proteins XP_001415359.1, XP_001415578.1, XP_001416135.1, XP_001418717.1, XP_001419043.1, XP_001419168.1, XP_001421374.1, XP_001422222.1, XP_001422692.1, XP_001422693.1, XP_001415368.1, XP_001415425.1, XP_001415474.1 and XP_001415497.1 were common to the first and the second lists. All of these identified proteins are false positives. Tables 3.4 and 3.6 show the counts of TPs, FPs and FNs with 100 PPM as the threshold value.

10	XP_001416177.1	4526	7	XP_001416378.1	18193	0.3333	YP_778511.1	231
7	XP_001415739.1	3600	7	XP_001416972.1	2378	0.2857	XP_001416438.1	80
7	XP_001416378.1	18193	6	XP_001421248.1	4434	0.2500	NP_683769.1	155
7	XP_001416972.1	2378	5	XP_001415739.1	3600	0.2500	NP_683794.1	43
6	XP_001421248.1	4434	5	XP_001416177.1	4526	0.2500	NP_817140.1	50
6	YP_778511.1	231	5	XP_001417211.1	874	0.2500	NP_817210.1	43
5	XP_001417211.1	874	5	XP_001418473.1	1770	0.2500	NP_817260.1	40
5	XP_001417790.1	1676	5	XP_001421657.1	743	0.2500	NP_817262.1	48
5	XP_001418473.1	1770	5	XP_001421977.1	1370	0.2500	XP_001416183.1	52
5	XP_001421657.1	743	5	XP_001422630.1	4395	0.2500	YP_001152065.1	50
5	XP_001421977.1	1370	4	NP_683861.1	1450	0.2500	YP_001152090.1	50
5	XP_001422630.1	4395	4	XP_001416600.1	737	0.2500	YP_001152098.1	51
4	NP_683861.1	1450	4	XP_001416917.1	1220	0.2500	YP_001152160.1	45
4	XP_001416600.1	737	4	XP_001417140.1	1186	0.2500	YP_636240.1	44
4	XP_001416917.1	1220	4	XP_001417525.1	1938	0.2500	YP_762308.1	52
4	XP_001417140.1	1186	4	XP_001418349.1	1199	0.2500	YP_784471.1	100
4	XP_001417525.1	1938	4	XP_001419308.1	3790	0.2000	NP_683849.1	121
4	XP_001418349.1	1199	4	XP_001420157.1	988	0.2000	NP_817204.2	62
4	XP_001419308.1	3790	4	XP_001420401.1	2320	0.2000	XP_001415824.1	63
4	XP_001420157.1	988	4	XP_001421119.1	1476	0.2000	XP_001417070.1	60
4	XP_001420401.1	2320	4	XP_001421255.1	3191	0.2000	XP_001417289.1	56
4	XP_001421119.1	1476	4	XP_001421282.1	2283	0.2000	XP_001417622.1	61
4	XP_001421255.1	3191	4	YP_001001592.1	1868	0.2000	XP_001419249.1	59
4	XP_001421282.1	2283	4	YP_001001595.1	2262	0.2000	XP_001419340.1	64
4	YP_001001592.1	1868	4	YP_001294246.1	2281	0.2000	XP_001422327.1	59
4	YP_001001595.1	2262	4	YP_778553.1	1974	0.2000	YP_001152142.1	55
4	YP_001294246.1	2281	3	NP_683769.1	155	0.2000	YP_001152178.1	55
4	YP_778553.1	1974	3	NP_683776.1	1373	0.2000	YP_001152207.1	65
3	NP_683769.1	155	3	XP_001415551.1	742	0.2000	YP_001152217.1	61
3	NP_683776.1	1373	3	XP_001415913.1	1281	0.2000	YP_001152226.1	58

Figure 3.30: Result from the pipeline when non-plant peptides from *Sus Scrofa* were used as input with 100 PPM as the threshold — test 4. The top 30 hits are displayed. Annotation for each column is the same as in Figure 3.21. In this case the pipeline reported NP_683769.1 in all the three lists. Proteins YP_778511.1 was common to the first and third lists. Furthermore, XP_001421657.1, XP_001416600.1 and XP_001420157.1 were reported in the first and the third lists.

10	XP_001416378.1	18193	6	XP_001416378.1	18193	0.3333	YP_636408.1	32
4	XP_001416972.1	2378	4	XP_001416972.1	2378	0.2500	NP_817140.1	50
3	XP_001415359.1	617	3	XP_001415359.1	617	0.2500	NP_817154.1	40
3	XP_001415578.1	357	3	XP_001415578.1	357	0.2500	YP_001152069.1	40
3	XP_001415955.1	1295	3	XP_001415955.1	1295	0.2500	YP_001152160.1	45
3	XP_001416135.1	797	3	XP_001416135.1	797	0.2500	YP_001152179.1	40
3	XP_001416982.1	3596	3	XP_001416982.1	3596	0.2500	YP_001152257.1	46
3	XP_001417092.1	3039	3	XP_001417092.1	3039	0.2500	YP_784384.1	100
3	XP_001417978.1	954	3	XP_001418717.1	680	0.2000	NP_817205.1	53
3	XP_001418717.1	680	3	XP_001418766.1	1479	0.2000	XP_001419101.1	56
3	XP_001418766.1	1479	3	XP_001419043.1	614	0.2000	XP_001419262.1	63
3	XP_001419043.1	614	3	XP_001419168.1	560	0.2000	XP_001420972.1	59
3	XP_001419168.1	560	3	XP_001420239.1	3608	0.2000	XP_001421828.1	60
3	XP_001419616.1	656	3	XP_001420313.1	258	0.2000	XP_001421864.1	64
3	XP_001420239.1	3608	3	XP_001421248.1	4434	0.2000	YP_001152177.1	64
3	XP_001420313.1	258	3	XP_001421374.1	678	0.1667	XP_001415598.1	68
3	XP_001421248.1	4434	3	XP_001422222.1	797	0.1667	XP_001419093.1	76
3	XP_001421374.1	678	3	XP_001422692.1	383	0.1667	XP_001419821.1	74
3	XP_001422222.1	797	3	XP_001422693.1	312	0.1667	XP_001420342.1	69
3	XP_001422241.1	577	3	YP_001123088.1	1791	0.1667	XP_001421240.1	78
3	XP_001422692.1	383	3	YP_538912.1	2278	0.1667	XP_001421641.1	78
3	XP_001422693.1	312	3	YP_636174.1	2178	0.1667	XP_001421916.1	70
3	YP_001123088.1	1791	3	YP_636177.1	3462	0.1667	YP_358599.1	66
3	YP_538912.1	2278	2	NP_084759.1	2434	0.1538	XP_001418170.1	168
3	YP_636174.1	2178	2	XP_001415368.1	484	0.1500	XP_001420313.1	258
3	YP_636177.1	3462	2	XP_001415425.1	654	0.1429	NP_683844.1	91
2	NP_084759.1	2434	2	XP_001415474.1	584	0.1429	XP_001416438.1	80
2	XP_001415368.1	484	2	XP_001415497.1	357	0.1429	XP_001417126.1	81
2	XP_001415425.1	654	2	XP_001415569.1	440	0.1429	XP_001418247.1	89
2	XP_001415474.1	584	2	XP_001415616.1	433	0.1429	XP_001419012.1	182
2	XP_001415497.1	357	2	XP_001415739.1	3600	0.1429	XP_001422736.1	85

Figure 3.31: Result from the pipeline when non-plant peptides from *Mus musculus* used as input with 100 PPM as the threshold value — test 5. Displayed are the top 30 hits. Annotation for each column is the same as in Figure 3.21. Protein XP_001420313.1 was reported in all the three potential plant protein lists. In total there were 14 proteins common to the first and second lists.

Further tests were performed where only known non-plant peptides from *Mus musculus* and *Sus scrofa* were used as input and varied threshold values used. The pipeline gave varied results. It was observed for cases where the tolerance/threshold value was 100 PPM (such as the two cases reported above), fewer proteins were observed in all the three reported lists by the pipeline.

In summary, during testing it was observed that the pipeline reported proteins consistently in at least one of the three potential lists when 4 or more peptides (per proteins) from known plant proteins were used as input with 100 PPM or 10 PPM selected as the threshold value. Table 3.4 shows a concise report of TPs for every test case considering the report of known plant in at least one of the three lists. Appendix Table A.1 shows the details of calculation of sensitivity and PPV for these cases and Table 3.5 shows a summary report of those calculations.

Table 3.6 provides a report on the counts of TPs, FPs and FNs for all the test cases using the rule mentioned earlier in the thesis for a "positive": that a protein must present in at least two of the three lists or preferably all the three lists. Appendix Table A.2 shows the details of calculations of sensitivity and PPV for the test cases. Finally, Table 3.7 gives a summary report with the counts of TPs, FPs, FNs along with information on sensitivity and PPV.

The ranking of the reported proteins within the lists is based on the decreasing order of total evidence (number of supporting peptides) found by the pipeline. With a stringent value (10 PPM), known plant proteins were positioned more towards the top of the lists. The positions changed when the threshold value was relaxed (100 PPM). This is an understandable behavior. A small window size limits the probability of many matches of a peptide sequence based on its m/z value. With a relaxed window size the probability of random matches increases, affecting the positions of the known plant proteins in the lists.

From Table 3.7, test case 1, it can be seen that 100 PPM as the threshold value showed better results in terms of sensitivity and PPV when compared with the other two threshold values. The likelihood of identifying TPs from among all actual positives was 66%. The odds are that, of the reported positives, 15% are true proteins (i.e. proteins actually present in the sample). A threshold value of 10 PPM showed identical sensitivity; however, the PPV was 8.6% compared to 15% with 100 PPM. Similarly in test case 2, compared to 10 and 50 PPM, 100 PPM showed better results. Again, threshold values of 100 and 10 PPM showed 100% sensitivity; however, the PPV with 100 PPM was 50% compared to 14.8% with 10 PPM.

From Table 3.7 it can be observed that using an equal number of peptides tends to increase the sensitivity and PPV for the pipeline. In case 1 unequal numbers of peptides are used (2, 11, and 11), while in test cases 2 and 3 four peptides are drawn for each protein. With the threshold values of 10 and 100PPM, having an equal number of peptides per protein gives better sensitivity and PPV values than having an unequal number. For example, for 100PPM, sensitivity goes from 66% in test case 1 to 100% in test cases 2 and 3. As another example, for 10PPM, PPV goes from 8.6% in test case 1 to 14.8% in test case 2 and 12.5% in test case 3. When the threshold value is 50PPM, however, having equal or unequal numbers of peptides per protein gives mixed results. As an example, sensivity goes from 33% in test case 1 to 50% in test case 2 (an increase), but to 0% in test case 3 (a decrease). However, in general with an equal number of peptides per protein in

the input data the pipeline was better able to report proteins.

In test case 4 and 5, only FPs were reported. From this report of FPs we can attempt to infer a rough estimate of the background level of FPs for test cases 1, 2 and 3. In test case 1 an unequal number of peptides per protein were used as input. Similarly, in test case 5, unequal number of peptides were used as input. The total number of peptides in test case 1 is almost the same as in test case 5. A significant proportion of FPs reported in test case 1 might be due to the background FP level. In test cases 2 and 4 equal numbers of peptides were used as input. About half of the FPs reported in test case 2 could be due to the background level of FPs. Again, in test cases 3 and 4, equal numbers of peptides per protein were used as input. In this case about half of the FPs reported in test case 3 may be due to the background noise.

In summary, the designed pipeline was able to identify proteins based on m/z values of the peptides in the input files. Plant proteins were identified and were among the top hits. However, other proteins (false positives) were also reported by the pipeline. From Table 3.7 it can be observed that test cases 2 and 3 showed better sensitivity and PPV. In test case 2 with 100 PPM, the sensitivity and PPV was highest compared to other test cases. About one fourth of the proteins were identified. It is expected that with real data set (all the non-plant peptide from a Pro Group report) the pipeline should be able to report approximately one fourth of proteins consistently across at least two of the three lists with 10 and 100 PPM as the threshold values.

Table 3.4: Results showing TPs reported in each of the three lists from five test cases. The second column shows the total count of plant peptides used as input followed by (in brackets) the total count of proteins from which those peptides were obtained. For example, 24 plant peptides were obtained from 3 different plant proteins in test case 1. The third column shows the corresponding information for non-plant peptides used as input. The fourth column of the table shows the number of peptides associated with different proteins used as input. For example, in test case 1, (2+11+11) signifies that 2 peptides were taken from one protein, and 11 each from two proteins. Similarly, in test case 3, 4 peptides each from 3 plant proteins (4∗3) and 4 peptides each from 3 non-plant proteins (4∗3) were selected, for a total of 24 ((4∗3)+(4∗3)) peptides used as input.

Test Case	Total Plant Peptides Used as Input (Total Proteins)	Total Non-Plant Peptides Used as Input (Total Proteins)	Number of Peptides Per Protein	Threshold Value in PPM	Total True Positives Within Top 30 (List 1)	Total True Positives Within Top 30 (List 2)	Total True Positives Within Top 30 (List 3)	Total Proteins Reported Consistently In the Three Lists
1	24 (3)	0	2+11+11	100	2	2	1	1
1	24 (3)	0	2+11+11	50	0	1	1	0
1	24 (3)	0	2+11+11	10	2	2	1	1
2	16 (4)	0	4+4+4+4	100	4	4	1	1
2	16 (4)	0	4+4+4+4	50	1	2	1	0
2	16 (4)	0	4+4+4+4	10	4	2	2	2
3	12 (4)	12 (4)	(4*3)+(4*3)	100	3	3	0	0
3	12 (4)	12 (4)	(4*3)+(4*3)	50	0	1	0	0
3	12 (4)	12 (4)	(4*3)+(4*3)	10	3	3	2	2
4	0	20 (4)	5+5+5+5	100	0	0	0	0
5	0	23 (3)	9+4+10	100	0	0	0	0

Table 3.5: Summary report for sensitivity and PPV for all the test cases. For the description of first five columns refer to Table 3.4. In the last two columns List1, List2 and List3 refers to the three potential lists from the pipeline. The calculation of sensitivity and PPV is shown in Appendix Table A.1. In the table entry "NC" indicated a value that was non-calculable. For example, in test case 4 only FPs are reported, therefore sensitivity cannot be calculated.

Test Case	**Total Plant Peptides Used as Input (Total Proteins)**	**Total Non-Plant Peptides Used as Input (Total Proteins)**	**Number of Peptides Per Proteins**	**Threshold Value in PPM**	**Sensitivity**	**PPV**
1	24(3)	0	2+11+11	100	List1: 66%	List1: 6.6%
					List2: 66%	List2: 6.6%
					List3: 33%	List3: 3.3%
				50	List1: 0%	List1: 0%
					List2: 33%	List2: 3.3%
					List3: 33%	List3: 3.3%
				10	List1: 66%	List1: 6.6%
					List2: 50%	List2: 6.6%
					List3: 50%	List3: 3.3%
2	16(4)	0	4+4+4+4	100	List1: 100%	List1: 13%
					List2: 100%	List2: 7%
					List3: 25%	List3: 3%
				50	List1: 25%	List1: 3%
					List2: 50%	List2: 6.6%
					List3: 25%	List3: 3%
				10	List1: 100%	List1: 13%
					List2: 100%	List2: 13%
					List3: 50%	List3: 6.6%
3	12(4)	12(4)	(4*3)+(4*3)	100	List1: 75%	List1: 10%
					List2: 75%	List2: 10%
					List3: 6.6%	List3: 6.6%
				50	List1: 0%	List1: 0%
					List2: 25%	List2: 3.3%
					List3: 0%	List3: 0%
				10	List1: 100%	List1: 10%
					List2: 100%	List2: 10%
					List3: 6.6%	List3: 6.6%
4	0	20(4)	(5+5+5+5)	100	List1: NC List2: NC List3: NC	List1: 0% List2: 0% List3: 0%
5	0	23(3)	(9+4+10)	100	List1: NC List2: NC List3: NC	List1: 0% List2: 0% List3: 0%

Table 3.6: Results showing identified proteins reported in at least two of the three lists from five test cases. The second column shows the total count of plant peptides used as input followed by (in brackets) the total count of proteins from which those peptides were obtained. For example, 24 plant peptides were obtained from 3 different plant proteins in test case 1. The third column shows the corresponding information for the non-plant peptides used as input. The fourth column of the table shows the number of peptides associated with different proteins used as input. For example, in test case 1, (2+11+11) signifies that 2 peptides were taken from one protein, and 11 each from two proteins. Similarly, in test case 3, 4 peptides each from 3 plant proteins (4∗3) and 4 peptides each from 3 non-plant proteins (4∗3) were selected, for a total of 24 ((4∗3)+(4∗3)) peptides used as input.

Test Case	Total Plant Peptides Used as Input (Total Proteins)	Total Non-Plant Peptides Used as Input (Total Proteins)	Number of Peptides Per Protein	Threshold Value in PPM	TP	FP	FN
1	24 (3)	0	2+11+11	100	2	11	1
1	24 (3)	0	2+11+11	50	1	16	2
1	24 (3)	0	2+11+11	10	2	21	1
2	16 (4)	0	4+4+4+4	100	4	8	0
2	16 (4)	0	4+4+4+4	50	2	10	2
2	16 (4)	0	4+4+4+4	10	4	23	0
3	12 (4)	12 (4)	(4*3)+(4*3)	100	3	15	0
3	12 (4)	12 (4)	(4*3)+(4*3)	50	0	23	3
3	12 (4)	12 (4)	(4*3)+(4*3)	10	3	21	0
4	0	20 (4)	5+5+5+5	100	0	6	0
5	0	23 (3)	9+4+10	100	0	15	0

Table 3.7: Summary report for TPs, FPs, FNs, sensitivity and PPV for all the test cases. For the description of first five columns refer to Table 3.4. The next three columns are described in Table 3.6. Appendix Table A.2 shows the calculations for sensitivity and PPV. In the table entry "NC" stands for non-calculable value. For example, in test case 4 TPs and FNs have zero values. Therefore sensitivity cannot be calculated.

Test Case	**Total Plant Peptides Used as Input (Total Proteins)**	**Total Non-Plant Peptides Used as Input (Total Proteins)**	**Number of Peptides Per Proteins**	**Threshold Value in PPM**	**TP**	**FP**	**FN**	**Sensitivity**	**PPV**
1	24(3)	0	2+11+11	100	2	11	1	66%	15%
				50	1	16	2	33%	5.8%
				10	2	21	1	66%	8.6%
2	16(4)	0	4+4+4+4	100	4	8	0	100%	50%
				50	2	10	2	50%	16.6%
				10	4	23	0	100%	14.8%
3	12(4)	12(4)	(4*3)+(4*3)	100	3	15	0	100%	16.6%
				50	0	23	0	NC	0%
				10	3	21	0	100%	12.5%
4	0	20(4)	(5+5+5+5)	100	0	6	0	NC	0%
5	0	23(3)	(9+4+10)	100	0	15	0	NC	0%

Chapter 4

Results

The main goal of the research presented here was to develop a bioinformatics pipeline that uses m/z values of non-plant peptides to recover plant proteins. The pipeline reported the output in the form of three potential plant protein lists. Proteins which were consistently reported in all three lists were more likely to be present in the sample. The methodology that was designed and the pipeline thus implemented allowed the use of the non-plant peptides to report plant proteins. Higher MW proteins reported by the pipeline should be regarded with caution due to the bias in the pipeline (see Section 3.4). Proteins either common to all the three lists or common to two of the three lists are highlighted. Such proteins are likely to be present in the sample.

Three separate experiments were conducted to evaluate the response of the pipeline when real data sets were used as input. Three reports from the Pro Group software were arbitrarily selected for the investigation. All the non-plant peptides from each of these reports were used as input. The total count of non-plant peptides were different in each of the reports. Two threshold values (100 and 10 PPM) were used in each experiment.

4.1 Experiment 1

In the first chosen Pro Group report there were a total of 140 proteins. 44 ($\approx 31\%$) were non-plant proteins. The non-plant proteins reported by Pro Group had different peptide counts. All the reported non-plant peptides, 50 in total, were used as input to the pipeline. Figure 4.1 shows the raw data. Three peptides lacked the needed charge state (see Section 3.3) and were excluded from the analysis. Thus 47 non-plant peptides were used by the pipeline for protein identification. Figure 4.2 shows all the processed peptides used by the pipeline with their associated values. Figure 4.3 shows the output from the pipeline using 100 PPM as the selected threshold value. Figure 4.5 shows a portion of input data followed by evidence (peptides) used by the pipeline to report plant proteins. No proteins were present in all three lists. There were a few proteins that were consistent across the first and second lists. However, as indicated by the lengths of the sequences, all such proteins are higher molecular weight proteins, and should not be accepted to represent proteins present in the sample.

ISGLIYEETR
TLUGFGG
TVTSMDVVYALJ
AGFAGDDAPR
TLADYNIQJ
AEFVEVTJ
ATEEQLJ
LVTDLTJ
YLYEIAR
APVLSDSSCJ
LSSPATLNSR
VATVSLPR
IAFSVSSAVDSR
QKEJ
EAYPGDIFYLHSR
IASIRVGESMLGR
TPLVSNLLAFLR
DGIDYAALLEHJJ
ELPKNVIVR
GDLAVLAAR
LAG
EIIK
EIPVEVIR
IYTJ
GDLEPLAAR
LVGMAHJADTPGR
AIEHEPGR
ESSSSTEGR
LSKDELTELJ
IDYGINUSSAJ
AVVFEGPSVQEANHR
YKAFEJPAK
YKAFEKPAJ
VATTANAAPTAAPK
IAALJ
VVDAIAJ
ASPVLLAMQR
FYAPJ
INDVJ
VTKGEFIGLANR
DNIQGITKPVIR
FLDJESER
DNIQGITJ
DMLEGVMEISEMRVR
MLVNFVIEK
IESGHAGGGAVDAR
EISPGSGPGEIR
CESAAHEADLJR
GHAGPGAGVRR
SLLLDAASPLFRER

Figure 4.1: Raw data used as input to the pipeline in experiment 1. A total of 50 non-plant peptides were used as input the pipeline.

531.3014	133.8325	4	3	QKEK
501.3160	168.1125	3	2	EIIK
514.3477	172.4564	3	2	IAALK
523.3004	175.4407	3	2	IYTK
587.3277	196.7831	3	2	INDVK
624.3270	209.1162	3	2	FYAPK
907.4508	227.8699	4	3	AIEHEPGR
714.4274	239.1497	3	2	VVDAIAK
1022.5029	256.6329	4	3	FLDKESER
1033.5526	259.3953	4	4	GHAGPGAGVRR
788.4640	263.8285	3	2	LVTDLTK
1066.6495	267.6696	4	3	ELPKNVIVR
1080.5964	271.1563	4	4	YKAFEKPAK
1080.5964	271.1563	4	4	YKAFEKPAK
817.4177	273.4798	3	2	ATEEQLK
841.5019	281.5078	3	2	VATVSLPR
1174.6439	294.6682	4	3	LSKDELTELK
884.5077	295.8431	3	2	GDLAVLAAR
887.4708	296.8308	3	2	DNIQGITK
921.4804	308.1673	3	2	AEFVEVTK
926.4859	309.8358	3	2	YLYEIAR
938.3937	313.8051	3	2	ESSSSTEGR
940.4974	314.5063	3	2	GDLEPLAAR
1295.6213	324.9125	4	3	IESGHAGGGAVDAR
975.4407	326.1541	3	2	AGFAGDDAPR
1303.7243	326.9383	4	3	VTKGEFIGLANR
1328.6138	333.1606	4	4	CESAAHEADLKR
1005.4797	336.1671	3	2	APVLSDSSCK
1351.7025	338.9328	4	4	LVGMAHKADTPGR
1352.7770	339.2014	4	3	DNIQGITKPVIR
1387.7600	347.9472	4	3	IASIRVGESMLGR
1044.5560	349.1925	3	2	LSSPATLNSR
1064.5498	355.8571	3	2	TLADYNIQK
1084.6059	362.5425	3	2	ASPVLLAMQR
1091.6045	364.8754	3	2	MLVNFVIEK
1471.7665	368.9488	4	4	DGIDYAALLEHKK
1566.7462	392.6937	4	3	EAYPGDIFYLHSR
1179.6130	394.2115	3	2	ISGLIYEETR
1586.8775	397.7266	4	3	SLLLDAASPLFRER
1197.5983	400.2066	3	2	EISPGSGPGEIR
1638.8109	410.7099	4	3	AVVFEGPSVQEANHR
1229.5924	410.8713	3	2	IDYGINYSSAK
1237.6300	413.5505	3	2	IAFSVSSAVDSR
1282.6876	428.5697	3	2	VATTANAAPTAAPK
1325.6896	442.9037	3	2	TVTSMDVVYALK
1342.7968	448.6061	3	2	TPLVSNLLAFLR
1793.8466	449.4688	4	3	DMLEGVMEISEMRVR

Figure 4.2: List of non-plant peptides used as input in Experiment 1. The figure shows 47 peptides used in Experiment 1. Data in the first column shows the predicted MW of the peptide sequences, data in the second column shows the predicted m/z values, data in the third column shows the predicted charge state, data in the fourth column shows the total count of basic AAs in the peptide sequences, and the data in the last column are the actual peptide sequences.

23	XP_001416378.1	18193	19	XP_001416378.1	18193	0.7500	YP_001001582.1	52
16	XP_001415739.1	3600	12	YP_001123785.1	1775	0.5714	XP_001418851.1	90
13	YP_001123785.1	1775	11	XP_001422420.1	1614	0.5000	XP_001416689.1	51
12	XP_001422420.1	1614	10	XP_001415739.1	3600	0.5000	XP_001416690.1	576
11	XP_001417362.1	4591	10	XP_001417100.1	1563	0.5000	YP_001152201.1	41
11	XP_001421774.1	3182	10	XP_001417362.1	4591	0.4444	XP_001418877.1	105
11	YP_001004245.1	2294	10	YP_001004245.1	2294	0.4286	YP_358557.1	80
10	XP_001417100.1	1563	10	YP_001123173.1	1794	0.4000	XP_001416061.1	55
10	YP_001123173.1	1794	10	YP_001294329.1	1882	0.4000	XP_001422493.1	59
10	YP_001294329.1	1882	9	XP_001417358.1	1407	0.3750	NP_683804.1	100
9	XP_001417358.1	1407	9	XP_001418911.1	1823	0.3750	YP_717251.1	100
9	XP_001417391.1	1291	9	XP_001421016.1	3018	0.3750	YP_784471.1	100
9	XP_001418911.1	1823	9	YP_001001592.1	1868	0.3333	XP_001416305.1	115
9	XP_001421016.1	3018	9	YP_636357.1	1877	0.3333	XP_001417236.1	68
9	YP_001001592.1	1868	9	YP_778549.1	1933	0.3333	XP_001417861.1	191
9	YP_636357.1	877	8	NP_862818.1	2287	0.3333	XP_001418060.1	76
9	YP_778549.1	1933	8	XP_001416537.1	1658	0.3333	XP_001418881.1	187
8	NP_862813.1	1828	8	XP_001417391.1	1291	0.3333	XP_001419955.1	71
8	NP_862818.1	2287	8	XP_001418571.1	4390	0.3333	XP_001419962.1	361
8	XP_001416537.1	1658	8	XP_001419308.1	3790	0.3333	XP_001420232.1	68
8	XP_001418571.1	4390	8	XP_001419980.1	1428	0.3333	XP_001420315.1	154
8	XP_001419308.1	3790	8	XP_001421908.1	1744	0.3333	YP_358599.1	66
8	XP_001419980.1	1428	8	XP_001422696.1	1120	0.3333	YP_636427.1	37
8	XP_001421908.1	1744	8	YP_001001595.1	2262	0.3333	YP_717208.1	37
8	XP_001422696.1	1120	8	YP_001294243.1	1859	0.3000	XP_001417584.1	125
8	YP_001001595.1	2262	8	YP_538908.1	1887	0.3000	XP_001419280.1	122
8	YP_001294243.1	1859	8	YP_636177.1	3462	0.3000	XP_001420357.1	121
8	YP_538908.1	1887	7	NP_862813.1	1828	0.3000	YP_025819.1	119
8	YP_636177.1	3462	7	XP_001416048.1	4962	0.2857	XP_001416163.1	81
7	NP_084688.1	1386	7	XP_001416870.1	1782	0.2857	XP_001418173.1	82

Figure 4.3: Result from the pipeline for experiment 1 with 100 PPM as a threshold value. Annotation for each column is the same as in Figure 3.21. From a total of 47 peptides used as input, no proteins were present in all three lists.

When the threshold parameter was changed to 10 PPM (Figure 4.4), 4 proteins (XP_001416282.1, XP_001416864.1, XP_001416305.1 and XP_001420726.1) were observed consistently across the three lists. Therefore there is higher confidence that these proteins are present in the sample. The following information was obtained from the database maintained by NCBI[1] for the 4 proteins (all from kingdom Viridiplantae):

XP_001416282.1: Plastid ribosomal protein L10, imported to chloroplast, large ribosomal subunit [*Ostreococcus lucimarinus* CCE9901]

XP_001416864.1: Predicted protein [*Ostreococcus lucimarinus* CCE9901]

XP_001416305.1: predicted protein [*Ostreococcus lucimarinus* CCE9901]

XP_001420726.1: predicted protein [*Ostreococcus lucimarinus* CCE9901]

The above results suggest the possibility of a protein from the species *Ostreococcus lucimarinus* (green algae) being present in the sample. Proteins XP_001416307.1, XP_001416064.1, and XP_001416688.1 were consistent across the first and second lists. Two proteins XP_001417464.1 and XP_001420612.1 were consistent across the first and third lists. As can be observed, the reported plant proteins were not from wheat. However, it is possible that some close homologues exist in wheat.

One possible explanation for the report of a protein different from wheat is that the reference database that was used had more sequences from this *Ostreococcus* species than wheat. Indeed, when the database was analyzed, from a total of 11013 protein sequences present in the database, 7707 contained sequences from *Ostreococcus* species; i.e. sequences from genus *Ostreococcus* were predominant in the reference database (NCBI-REFSEQ). Other analysis revealed that the top hit in the initial and second lists of Figure 4.3) – i.e. the protein with ID XP_001416378.1 – had the largest AA count (18193). This implies that there would be more chances of non-plant peptides matching to this protein than to the other proteins present in the database. Such instances have already been discussed in Section 3.5 of Chapter 3.

[1] http://www.ncbi.nlm.nih.gov/. Last accessed November 22, 2008.

6	XP_001415739.1	3600	5	XP_001418911.1	1823	0.4000	XP_001422493.1	59
6	XP_001417790.1	1676	3	XP_001417790.1	1676	0.3333	XP_001417236.1	68
5	XP_001418911.1	1823	3	XP_001418094.1	1594	0.2500	XP_001420501.1	52
5	XP_001420612.1	381	3	XP_001418886.1	547	0.2222	**XP_001416305.1**	115
4	XP_001416443.1	863	3	XP_001419308.1	3790	0.2000	YP_001152241.1	53
4	XP_001417464.1	383	3	**XP_001420726.1**	211	0.1818	XP_001419489.1	140
3	XP_001418094.1	1594	3	XP_001421522.1	3454	0.1765	**XP_001420726.1**	211
3	XP_001418886.1	547	3	XP_001422369.1	1091	0.1667	NP_817179.1	69
3	XP_001419308.1	3790	3	XP_001422859.1	778	0.1667	XP_001420232.1	68
3	**XP_001420726.1**	211	3	YP_001123785.1	1775	0.1667	XP_001420612.1	381
3	XP_001421522.1	3454	3	YP_778553.1	1974	0.1429	NP_084763.1	87
3	XP_001422369.1	1091	2	NP_084773.1	2280	0.1429	**XP_001416864.1**	170
3	XP_001422859.1	778	2	NP_862739.1	507	0.1429	XP_001419246.1	89
3	YP_001123785.1	1775	2	NP_862818.1	2287	0.1429	XP_001419454.1	180
3	YP_778553.1	1974	2	XP_001416064.1	381	0.1429	YP_001123097.1	88
2	NP_084773.1	2280	2	**XP_001416282.1**	212	0.1429	YP_001123784.1	88
2	NP_862739.1	507	2	**XP_001416305.1**	115	0.1429	YP_358607.1	81
2	NP_862818.1	2287	2	XP_001416307.1	486	0.1429	YP_717235.1	87
2	XP_001415990.1	663	2	XP_001416378.1	18193	0.1333	XP_001417464.1	383
2	XP_001416027.1	434	2	XP_001416443.1	863	0.1333	XP_001418881.1	187
2	XP_001416064.1	381	2	XP_001416598.1	536	0.1333	XP_001419022.1	185
2	XP_001416197.1	554	2	XP_001416688.1	260	0.1333	XP_001419754.1	184
2	**XP_001416282.1**	212	2	XP_001416798.1	637	0.1333	XP_001420674.1	189
2	**XP_001416305.1**	115	2	**XP_001416864.1**	170	0.1250	XP_001421182.1	92
2	XP_001416307.1	486	2	XP_001417017.1	1393	0.1250	XP_001421496.1	95
2	XP_001416378.1	18193	2	XP_001417123.1	302	0.1250	XP_001422008.1	102
2	XP_001416598.1	536	2	XP_001417144.1	233	0.1250	YP_001294225.1	92
2	XP_001416688.1	260	2	XP_001417230.1	898	0.1250	YP_358568.1	98
2	XP_001416798.1	637	2	XP_001417304.1	4003	0.1250	YP_358581.1	103
2	**XP_001416864.1**	170	2	XP_001417362.1	4591	0.1176	**XP_001416282.1**	212

Figure 4.4: Result from the pipeline for experiment 1 with 10 PPM as the threshold value. Annotation for each column is the same as in Figure 3.21. A total of 47 peptides were used as input. The proteins are color-coded according to their appearance in the three lists. Four proteins (XP_001420726.1, XP_001416305.1, XP_001416282.1 and XP_001416864.1) color-coded (in brown, light-indigo, dark-green and indigo, respectively) and in bold were observed consistently across the three lists. Three proteins (XP_001418886.1, XP_001416307.1 and XP_001416064.1) color-coded (light-brown, blue and orange, respectively) were observed consistently across first and second lists, and two proteins, XP_001417464.1 and XP_001420612.1 (color-coded in red and light-blue, respectively) were observed consistently across first and third lists.

Non-plant peptides in the input data:

587.3277 196.7831 3 2 INDVK

1022.5029 256.6329 4 3 FLDKESER

1174.6439 294.6682 4 3 LSKDELTELK

Evidence for protein XP_001420726.1:

171 176 587.3276 196.7831 3 2 XP_001420726.1 DIVGGK

207 210 511.2502 256.6323 2 3 XP_001420726.1 EHAR

171 176 587.3276 294.6710 2 2 XP_001420726.1 DIVGGK

hits.196.7831

hits.256.6329

hits.294.6682

Evidence for protein XP_001416864.1:

9 14 587.3276 196.7831 3 2 XP_001416864.1 DVGLGK

9 14 587.3276 294.6710 2 2 XP_001416864.1 DVGLGK

hits.196.7831

hits.294.6682

Evidence for protein XP_001416282.1:

201 206 587.3276 196.7831 3 2 XP_001416282.1 AIAEGK

201 206 587.3276 294.6710 2 2 XP_001416282.1 AIAEGK

hits.196.7831

hits.294.6682

Evidence for protein XP_001416305.1:

73 77 587.3276 196.7831 3 2 XP_001416305.1 EQALK

73 77 587.3276 294.6710 2 2 XP_001416305.1 EQALK

hits.196.7831

hits.294.6682

Figure 4.5: Evidence for the proteins reported by the pipeline — experiment 1, 100 PPM. The top of the figure shows the peptides used by the pipeline for protein identification. From left to right the values are MW, m/z value, charge state, basic AA count, and the peptide sequence. The evidence for the four proteins is shown in the remainder of the figure. The columns constituting that information are (from left to right) start and end location of the peptide in the actual sequence, followed by MW, m/z value, charge state, basic AA count, the protein ID, and the peptide sequence. The captured peptides falling in the interval based on the m/z value is stored in intermediate files (hits.XXX, where XXX is the m/z value of the peptide of the input sequence to the pipeline) generated by the pipeline.

4.2 Experiment 2

In this case, of a total of 236 proteins reported by Pro Group, 65 ($\approx$ 27%) were non-plant proteins. The non-plant proteins reported by Pro Group had a range of peptide counts. All the reported non-plant peptides, 80 in total, were used as input to the pipeline of which 79 were used by the pipeline for protein identification. One peptide did not met the requirement of charge state. With 100 PPM as the threshold value, one protein XP_001416827.1 (predicted protein *Ostreococcus lucimarinus* CCE9901) was reported across the three potential plant protein lists (Figure 4.6). Protein XP_001419660.1 (predicted protein *Ostreococcus lucimarinus* CCE9901) was common to the first and second lists. However, when 10 PPM was used as the threshold value, 2 proteins, XP_001015435.1 (predicted protein *Ostreococcus lucimarinus* CCE9901) and XP_001421496.1 (P-ATPase family transporter: calcium ion, *Ostreococcus lucimarinus* CCE9901), were reported in all the three plant lists (Figure 4.7). Two proteins, XP_001420501.1 (predicted protein *Ostreococcus lucimarinus* CCE9901) and XP_001417211.1 (predicted protein *Ostreococcus lucimarinus* CCE9901) were reported at least in the two (first and third, second and third respectively) potential plant lists.

4.3 Experiment 3

The third chosen Pro Group report contained a total of 236 proteins of which 70 ($\approx$ 29%) were non-plant proteins. The reported non-plant proteins reported had a variety of raw peptide counts. All the reported non-plant peptides, 121 in total, were collected from the Pro Group report and were used as input to the pipeline. After processing, 114 non-plant peptides were used by the pipeline for protein identification. Seven peptides either failed to meet the requirement of charge state or molecular weight. Figure 4.8 shows that two proteins, XP_001416519.1 (predicted protein *Ostreococcus lucimarinus* CCE9901) and XP_001417105.1 (predicted protein *Ostreococcus lucimarinus* CCE9901), were reported consistently in the first and second lists with a threshold value of 100 PPM. However, when the threshold values was changed to 10 PPM a total of 11 proteins, XP_001422702.1 (Photosystem I light harvesting complex, chlorophyll a/b binding *Ostreococcus lucimarinus* CCE9901), XP_001415522.1 (predicted protein *Ostreococcus lucimarinus* CCE9901), XP_001415834.1 (lysine decarboxylase-related protein *Ostreococcus lucimarinus* CCE9901), XP_001417611.1 (predicted protein *Ostreococcus lucimarinus* CCE9901), XP_001418234.1 (predicted protein *Ostreococcus lucimarinus* CCE9901), XP_001418787.1 (Tic110 family transporter: chloroplast inner envelope protein Tic110 *Ostreococcus lucimarinus* CCE9901), XP_001419632.1 (predicted protein *Ostreococcus lucimarinus* CCE9901), XP_001420622.1 (predicted protein *Ostreococcus lucimarinus* CCE9901), XP_001420971.1 (predicted protein *Ostreococcus lucimarinus* CCE9901), XP_001421479.1 (pre-

dicted protein *Ostreococcus lucimarinus* CCE9901), XP_001421664.1(predicted protein *Ostreococcus lucimarinus* CCE9901) were observed consistently across the two potential plant lists (Figure 4.9).

42	XP_001416378.1	18193	27	XP_001416378.1	18193	1.0000	XP_001420501.1	52
25	XP_001415739.1	3600	18	XP_001416972.1	2378	0.6667	XP_001419955.1	71
23	XP_001416177.1	4526	14	XP_001415739.1	3600	0.5714	YP_874709.1	90
22	XP_001416972.1	2378	14	YP_636192.1	2596	0.5556	XP_001419514.1	115
18	XP_001420239.1	3608	13	XP_001416048.1	4962	0.5000	NP_817281.1	41
15	XP_001417215.1	1345	11	XP_001422630.1	4395	0.5000	XP_001415423.1	51
15	YP_636192.1	2596	10	XP_001416177.1	4526	0.5000	YP_001001582.1	52
14	XP_001416048.1	4962	10	XP_001417362.1	4591	0.5000	YP_001152095.1	67
12	XP_001422630.1	4395	10	YP_001001595.1	2262	0.5000	YP_001152192.1	50
11	XP_001417362.1	4591	9	XP_001417140.1	1186	0.5000	YP_001152236.1	45
11	XP_001419660.1	755	9	XP_001417293.1	2759	0.5000	YP_636347.1	52
11	YP_001123258.1	1778	9	XP_001419660.1	755	0.4444	XP_001417146.1	105
11	YP_636177.1	3462	9	YP_001123258.1	1778	0.4286	YP_001001516.1	86
10	**XP_001416827.1**	369	9	YP_358645.1	2290	0.4286	YP_717235.1	87
10	XP_001417140.1	1186	9	YP_636177.1	3462	0.4000	XP_001415405.1	118
10	XP_001417293.1	2759	9	YP_778553.1	1974	0.4000	XP_001416383.1	64
10	XP_001418571.1	4390	8	XP_001415633.1	2197	0.4000	XP_001420299.1	60
10	XP_001419271.1	604	8	XP_001417440.1	1027	0.4000	XP_001420865.1	55
10	XP_001420985.1	467	8	XP_001418259.1	2024	0.4000	XP_001421521.1	61
10	XP_001421296.1	2136	8	XP_001418331.1	1870	0.4000	YP_001152191.1	56
10	XP_001421774.1	3182	8	XP_001418571.1	4390	0.4000	YP_001152205.1	58
10	XP_001421977.1	1370	8	XP_001420239.1	3608	0.3750	XP_001421496.1	95
10	XP_001422420.1	1614	8	XP_001421264.1	708	0.3750	YP_636451.1	100
10	YP_001001595.1	2262	8	XP_001421977.1	1370	0.3750	YP_717251.1	100
10	YP_001294329.1	1882	8	YP_001001592.1	1868	0.3529	XP_001421633.1	220
10	YP_358645.1	2290	8	YP_001294329.1	1882	0.3448	**XP_001416827.1**	369
10	YP_778553.1	1974	8	YP_001294415.1	2260	0.3333	NP_817145.1	68
9	XP_001416088.1	2253	7	NP_862813.1	1828	0.3333	XP_001415483.1	146
9	XP_001417304.1	4003	7	NP_862818.1	2287	0.3333	XP_001415833.1	145
9	XP_001417440.1	1027	7	XP_001415426.1	1546	0.3333	XP_001416152.1	35
9	XP_001417825.1	1153	7	XP_001416088.1	2253	0.3333	XP_001416262.1	106
9	XP_001418259.1	2024	7	**XP_001416827.1**	369	0.3333	XP_001417059.1	183
9	XP_001418331.1	1870	7	XP_001417211.1	874	0.3333	XP_001418105.1	78
9	XP_001421016.1	3018	7	XP_001417391.1	1291	0.3333	XP_001419213.1	108

Figure 4.6: Result from the pipeline for experiment 2 with 100 PPM as threshold value to the pipeline. A total of 80 non-plant peptides were used as input. Annotation for each column is the same as in Figure 3.21. Protein XP_001416827.1 (color-coded in indigo) was reported consistently across the three potential plant list. Protein XP_001419660.1 was common to the first and second lists.

13	XP_001417215.1	1345	5	XP_001416048.1	4962	1.0000	XP_001420501.1	52
6	XP_001416048.1	4962	4	YP_636192.1	2596	0.4000	YP_001152205.1	58
6	XP_001418911.1	1823	3	XP_001416296.1	1280	0.3750	**XP_001421496.1**	95
5	XP_001416378.1	18193	3	XP_001416378.1	18193	0.3333	YP_001294304.1	37
5	XP_001418450.1	1701	3	XP_001417211.1	874	0.3333	YP_636400.1	39
5	XP_001421646.1	560	3	XP_001417493.1	1135	0.2857	NP_904239.1	81
4	XP_001417304.1	4003	3	XP_001418450.1	1701	0.2857	XP_001419402.1	88
4	XP_001420501.1	52	3	XP_001418751.1	1418	0.2500	YP_001152158.1	48
4	XP_001421699.1	645	3	XP_001418911.1	1823	0.2308	XP_001418170.1	168
4	XP_001422420.1	1614	3	XP_001420239.1	3608	0.2222	XP_001419514.1	115
4	YP_636192.1	2596	3	**XP_001421496.1**	95	0.2222	XP_001419556.1	110
3	NP_084773.1	2280	3	XP_001421646.1	560	0.2222	YP_025759.1	111
3	**XP_001415435.1**	223	3	XP_001422369.1	1091	0.2000	NP_084760.1	54
3	XP_001416295.1	1051	3	XP_001422420.1	1614	0.2000	XP_001418789.1	123
3	XP_001416296.1	1280	3	XP_001422699.1	861	0.2000	XP_001419638.1	119
3	XP_001416870.1	1782	3	YP_001001595.1	2262	0.2000	XP_001420788.1	126
3	XP_001417056.1	991	2	NP_084773.1	2280	0.2000	XP_001420821.1	358
3	XP_001417211.1	874	2	**XP_001415435.1**	223	0.2000	XP_001422464.1	121
3	XP_001417493.1	1135	2	XP_001415603.1	925	0.2000	YP_025751.1	64
3	XP_001417762.1	611	2	XP_001415901.1	200	0.1818	XP_001420067.1	134
3	XP_001417978.1	954	2	XP_001415907.1	365	0.1818	XP_001420298.1	137
3	XP_001418147.1	596	2	XP_001416037.1	552	0.1667	**XP_001415435.1**	223
3	XP_001418170.1	168	2	XP_001416081.1	664	0.1667	XP_001415645.1	74
3	XP_001418654.1	293	2	XP_001416088.1	2253	0.1667	XP_001416840.1	155
3	XP_001418751.1	1418	2	XP_001416284.1	1545	0.1667	XP_001417613.1	147
3	XP_001419339.1	563	2	XP_001416295.1	1051	0.1667	XP_001418105.1	78
3	XP_001419629.1	412	2	XP_001416688.1	260	0.1667	XP_001418112.1	534
3	XP_001420239.1	3608	2	XP_001416790.1	163	0.1667	XP_001418421.1	70
3	XP_001421421.1	401	2	XP_001416818.1	429	0.1667	XP_001418862.1	145
3	**XP_001421496.1**	95	2	XP_001416870.1	1782	0.1667	XP_001420232.1	68

Figure 4.7: Result from the pipeline for experiment 2 with 10 PPM as threshold value. The figures shows the result when a total of 80 non-plant peptides were used as input to the pipeline. Annotation for each column is the same as in Figure 3.21. Protein XP_001415435.1 (color-coded in blue) and XP_001421496.1 (color-coded in dark blue) were reported in all three potential plant lists. Two other proteins, XP_001417211.1 and XP_001420501.1, were reported in at least two potential plant lists and color-coded in green and red, respectively.

51	XP_001416378.1	18193	40	XP_001416378.1	18193	0.6667	YP_001123841.1	33
23	XP_001418571.1	4390	23	XP_001418571.1	4390	0.5556	YP_717278.1	114
21	XP_001416177.1	4526	19	XP_001417362.1	4591	0.5000	NP_817177.1	43
20	XP_001415739.1	3600	17	XP_001421248.1	4434	0.5000	XP_001416952.1	72
19	XP_001417362.1	4591	15	XP_001416048.1	4962	0.5000	XP_001418060.1	76
17	XP_001421248.1	4434	14	XP_001421282.1	2283	0.5000	YP_001152095.1	67
15	XP_001416048.1	4962	13	XP_001415739.1	3600	0.5000	YP_001152134.1	41
15	XP_001420239.1	3608	13	XP_001416982.1	3596	0.5000	YP_001152181.1	43
14	XP_001421282.1	2283	12	XP_001416972.1	2378	0.5000	YP_001152253.1	40
13	XP_001416982.1	3596	12	XP_001417304.1	4003	0.4286	NP_689364.1	91
13	XP_001417741.1	1242	12	XP_001418754.1	2823	0.4000	NP_817226.1	122
12	XP_001416972.1	2378	12	XP_001418911.1	1823	0.4000	XP_001417014.1	63
12	XP_001417304.1	4003	12	XP_001422461.1	1283	0.4000	XP_001418884.1	55
12	XP_001418754.1	2823	12	YP_636177.1	3462	0.4000	XP_001422340.1	64
12	XP_001418911.1	1823	11	XP_001417741.1	1242	0.4000	XP_001422600.1	57
12	XP_001422461.1	1283	11	XP_001420717.1	2267	0.4000	YP_636220.1	125
12	YP_001294314.1	2271	11	XP_001421522.1	3454	0.3750	NP_683804.1	100
12	YP_636177.1	3462	11	XP_001422630.1	4395	0.3750	XP_001419460.1	104
11	NP_084759.1	2434	11	YP_001294314.1	2271	0.3750	YP_636181.1	92
11	XP_001416991.1	1362	10	NP_683831.1	1627	0.3333	NP_084677.1	111
11	XP_001417825.1	1153	10	XP_001415386.1	1271	0.3333	NP_817198.1	69
11	XP_001420618.1	1899	10	XP_001416519.1	808	0.3333	NP_862764.1	36
11	XP_001420717.1	2267	10	XP_001416870.1	1782	0.3333	XP_001418086.1	77
11	XP_001421522.1	3454	10	XP_001417105.1	864	0.3333	XP_001418912.1	75
11	XP_001422630.1	4395	10	XP_001420487.1	2198	0.3333	XP_001419946.1	70
10	NP_683831.1	1627	10	XP_001421119.1	1476	0.3333	XP_001420125.1	184
10	XP_001415386.1	1271	10	XP_001422408.1	873	0.3333	XP_001422318.1	76
10	XP_001416093.1	1395	9	NP_084759.1	2434	0.3333	XP_001422323.1	74
10	XP_001416519.1	808	9	NP_862818.1	2287	0.3333	XP_001422496.1	117
10	XP_001416870.1	1782	9	XP_001416093.1	1395	0.3333	YP_001122965.1	37
10	XP_001417105.1	864	9	XP_001416177.1	4526	0.3333	YP_001294386.1	37
10	XP_001417790.1	1676	9	XP_001416502.1	1189	0.3333	YP_024356.1	34
10	XP_001419639.1	740	9	XP_001417092.1	3039	0.3333	YP_636449.1	184
10	XP_001420487.1	2198	9	XP_001417740.1	2272	0.3333	YP_740392.1	105

Figure 4.8: Result from the pipeline for experiment 3 with 100 PPM as threshold value. Three lists showing the output from the pipeline when a total of 121 non-plant peptides were used as input to the pipeline. Annotation for each column is the same as in Figure 3.21. Two proteins, XP_001416519.1 and XP_001417105.1 (color-coded in red and navy blue, respectively), were reported consistently across two potential plant lists.

5	XP_001416378.1	18193	4	XP_001415684.1	1076	0.2857	YP_025821.1	83
5	YP_001123505.1	2293	4	XP_001415739.1	3600	0.2500	XP_001415423.1	51
5	YP_001123527.1	2293	4	XP_001416378.1	18193	0.2500	XP_001420501.1	52
4	XP_001415684.1	1076	4	XP_001422702.1	901	0.2500	XP_001422105.1	101
4	XP_001415739.1	3600	3	XP_001415522.1	506	0.2500	YP_001152201.1	41
4	XP_001417790.1	1676	3	XP_001416474.1	325	0.2000	NP_084760.1	54
4	XP_001422702.1	901	3	XP_001417362.1	4591	0.2000	XP_001418884.1	55
3	XP_001415522.1	506	3	XP_001417611.1	900	0.2000	XP_001419280.1	122
3	XP_001415834.1	928	3	XP_001418234.1	1103	0.2000	XP_001422327.1	59
3	XP_001416088.1	2253	3	XP_001418571.1	4390	0.2000	XP_001422340.1	64
3	XP_001416474.1	325	3	XP_001418787.1	901	0.2000	YP_001152118.1	56
3	XP_001417362.1	4591	3	XP_001419632.1	895	0.2000	YP_001152185.1	63
3	XP_001417611.1	900	3	XP_001420622.1	815	0.2000	YP_001152202.1	59
3	XP_001418200.1	376	3	XP_001420971.1	472	0.1667	NP_817198.1	69
3	XP_001418234.1	1103	3	XP_001421479.1	689	0.1667	XP_001416952.1	72
3	XP_001418392.1	852	3	XP_001421664.1	486	0.1667	XP_001418086.1	77
3	XP_001418571.1	4390	3	YP_001123505.1	2293	0.1667	XP_001419426.1	77
3	XP_001418787.1	901	3	YP_001123527.1	2293	0.1667	XP_001420232.1	68
3	XP_001418815.1	1599	3	YP_001294415.1	2260	0.1667	XP_001421202.1	68
3	XP_001419308.1	3790	3	YP_778549.1	1933	0.1429	NP_817237.1	91
3	XP_001419632.1	895	2	NP_084773.1	2280	0.1429	NP_817287.1	88
3	XP_001420161.1	617	2	NP_683774.1	1070	0.1429	XP_001418725.1	90
3	XP_001420622.1	815	2	XP_001415520.1	1361	0.1429	XP_001419402.1	88
3	XP_001420879.1	1177	2	XP_001415577.1	283	0.1333	XP_001419602.1	186
3	XP_001420971.1	472	2	XP_001415599.1	1503	0.1333	XP_001420887.1	189
3	XP_001421203.1	1089	2	XP_001415634.1	624	0.1250	NP_689346.1	104
3	XP_001421479.1	689	2	XP_001415637.1	524	0.1250	XP_001416522.1	103
3	XP_001421654.1	360	2	XP_001415802.1	842	0.1250	XP_001419379.1	100
3	XP_001421664.1	486	2	XP_001415828.1	897	0.1250	XP_001421081.1	206
3	XP_001421967.1	642	2	XP_001415834.1	928	0.1250	XP_001421182.1	92

Figure 4.9: Result from the pipeline for experiment 3 with 10 PPM as threshold value. The three lists shows output proteins when a total of 121 non-plant peptides were used as input to the pipeline. Annotation for each column is the same as in Figure 3.21. A total of 11 proteins (colour-coded in different colors) were reported consistently across the first two potential plant lists. Protein IDs are identified in the text above.

From the above experiments it was observed that more proteins were common across the three potential plant lists with 10 PPM as the selected threshold value in the three experiments. Also, with 10 PPM more proteins were common to at least two of the three lists (Table 4.1).

Table 4.1: Concise report from all the three experiments.

Experiment	**Peptides Used By The Pipeline**	**Threshold Value In PPM**	**Proteins Common To Two Lists**	**Proteins Common To All The Three Lists**
1	47	100	0	0
		10	5	4
2	79	100	1	1
		10	2	2
3	114	100	2	0
		10	11	0

Chapter 5

Discussion

5.1 Discussion of the Results

During testing of the pipeline (Section 3.5) with the known plant peptides and mixed peptides (plant and known non-plant) 100 PPM and 10 PPM threshold values showed better results, i.e. more proteins were reported common to the three lists. It was expected that with the real data sets (peptides reported by the Pro Group software) more plant proteins will be reported consistently in at least two lists with 10 and 100 PPM as the threshold values. From the experiments in Chapter 4 it was observed that with 10 PPM as the selected threshold value better results were obtained. The number of proteins reported across the three lists was higher than with 10 PPM as the threshold value. The number of proteins reported across the three lists was higher than with 10 PPM as the threshold value. The MS/MS analysis was completed at the UVIC Genome BC Proteomics Center. Their instrument had an accuracy of 25 PPM, and so the m/z values of the non-plant peptides used in this work had an accuracy of 25 PPM. More experiments would be needed to investigate the relationship (if any) between the report of higher number proteins common to the three lists with 10 PPM as the threshold value and the accuracy of the mass spectra (25 PPM). It is conceivable that a less stringent value, such as 50 or 100 PPM would lead to fewer proteins common across two or more lists since the more relaxed threshold would allow more proteins in each list. This would lower the probability that a particular protein would be among the top 30 hits in a particular list.

In Chapter 4, it was observed that the database contained more proteins from genus *Osterococcus* than from wheat or any other plant species. In order to obtain proteins sequences from wheat only, an alternative is to use a wheat-only database. However, there are certain limitations associated with this approach. Even though great advancements have been made in the past years in the field of plant proteomics, a catalogue of all the protein sequences in the proteome of an economical model species, such as wheat, still remains illusive [21]. Many initiatives have been taken to achieve this goal by 2010 [41].

The REFSEQ protein sequence database from NCBI was used in this thesis. Another option is to use multiple databases, such as NCBI-REFSEQ and Swiss-Prot [38], for protein identification. The protein sequences from the two databases can be filtered and processed separately as mentioned

earlier (Section 3.2.1.1). The pipeline will then correlate the m/z values of all the non-plant peptides with the m/z values of the peptides present in each of the databases. The proteins thus obtained would be reported to the user. The report of identical proteins from two different databases would raise the likelihood of the protein being present in the sample. One limitation of this approach is that the same protein sequence might be present in the two databases but with different identifiers. Then, since different protein identifiers were reported, it might not be immediately evident that the same protein is being reported. The use of two databases may also assist in cases where neither database has complete sequence information (e.g. for a particular species such as wheat). Then the union of the lists returned by each search could be used. Such an approach might be useful if a protein present in the test sample is only present in one of the databases.

5.2 Motivation to Use Mass-to-Charge Ratios of Non-Plant Peptides Instead of Homology-based Approach

In order to address the problem of reported non-plant "hits", molecular biologists typically use a sequence-based method for plant protein identification. The method relies on sequence alignment tools and databases for protein identication. Non-plant protein sequences are used as input to the sequence alignment tools, while the search is against a plant-only database. By default such alignment tools use evolutionarily based substitution matrices (for example PAM, BLOSUM) to find (potential) homologs of the query proteins. Unfortunately, AA substitutions which might be allowed by evolution do not necessarily conserve properties related to the MS/MS spectra.

Two experiments were conducted to test the efficacy of the sequence homology-based approach. Two arbitrarily selected non-plant proteins identified by Pro Group were selected. Using their accession numbers, the complete protein sequences were downloaded from the SwissProt[1] database. Each protein sequence was then given as the query to a BLAST[2] search of the SwissProt database restricted to green plants. Each search returned many "hits." Hits with the lowest expectation values were used for further analysis.

5.2.1 Experiments

Experiment 1

The following non-plant protein was reported by Pro Group: "AY513239 NID: - *Homo sapiens*, Accession Number AAR89906". A BLAST search returned the hit *XP_002499844.1, alpha-2 macroglobulin family-like protein from Micromonas* sp. RCC299 with 27% identity and an E-value of 4e-59. 6 other hits were reported. The two sequences AAR89906 and XP_002499844.1 were

[1] http://ca.expasy.org/databases.html. The database was last accessed on September 2006.
[2] http://ca.expasy.org/tools/blast/

digested using MS-Digest (version 4.0.7 available through ProteinProspector[3]) [5] using default settings [10, 11] for most of the parameters. Table 5.1 lists the non-default parameters. The spectrum was sorted by m/z ratio and than the values were manually compared for correspondence between the m/z values of the reported proteins (Table 5.2). The table only shows the top 18 sequences and corresponding m/z values. Manual comparison shows no significant matches between the theoretical spectra. That is, peaks in the original mass spectra which led to the identification of human protein AAR89906 would not have led to the identification of *Micromonas* sp. RCC 299. Inferring the presence of protein "alpha-2 macroglobulin family-like protein" in the sample based on the report of AAR89906 is not justified.

Table 5.1: User-defined parameters used to obtain the m/z values.

Parameter	Value
Database	UserProtein
Digest	Trypsin
Max. number of missed cleavages	2
Cys modified by	carbamidomethylation
Instrument	Q-Star

Table 5.2: Comparison of the m/z values obtained for the non-plant protein (reported by Pro Group) and the plant protein (reported by BLAST) from from experiment 1.

Non-Plant Protein Reported By Pro Group		Plant Protein Reported By BLAST	
m/z	Sequence	*m/z*	Sequence
804.4363	DSRWLK	818.4003	VDADDRK
806.4883	LKNIYR	846.5043	RLTSELK
815.4985	GIQELKK	855.3843	QSSDGSFK
818.4618	SLSVVDAK	856.5363	GLKPRSAK
820.3836	WADEATK	873.5040	IEGEVKAK
828.4938	KLPLDSR	874.5080	WLLMRR
832.4312	TIQEWR	887.5197	VIAATSPTK
854.4367	NLTTSYR	889.4738	GDGLTAVTR
856.3393	CETIFED	934.4662	ATEVNACVK
874.4350	QFLSSHR	944.4836	IEPEGFPR
890.5570	SVFLRLR	946.5932	SAKVTLTVK
891.4757	MLDKAWK	947.5309	GGLFLIDGR
902.4254	YSYDNLK	962.4836	GCDLATRAR
914.3448	CETIFED	965.4244	ADLMDSADK
935.5673	VLKLYGSR	974.5309	LTSELKQR
936.4754	RMDLMGNK	990.5214	ATAVTEAATR
943.5221	NWNVVRR	992.4717	ATEVNACVK
943.5432	RDDLLRR	1020.4891	GCDLATRAR

[3]http://prospector.ucsf.edu/prospector/4.0.7/html/msdigest.htm

Experiment 2

The following non-plant protein was reported by Pro Group: "A54324: carboxypeptidase H - American goosefish". A BLAST search returned a protein with accession number XP_001754759.1, a predicted protein from *Physcomitrella patens* subsp. *patens* as the best hit with 41% precent identity and an E-value of 1e-78 . There were 10 other hits returned. The same procedure was followed as mentioned above to match the mass spectra. Table 5.3 only shows the top 18 sequences and corresponding m/z values. Again little, if any, correspondence is evident. Based on sequence homology presence of *Physcomitrella patens* subsp. *patens* in the sample cannot be justified.

Table 5.3: Comparison of the m/z values obtained for the non-plant protein (reported by Pro Group) and the plant protein (reported by BLAST) from experiment 2.

Non-Plant Protein Reported By Pro Group		Plant Protein Reported By BLAST	
m/z	Sequence	m/z	Sequence
811.3369	DHDYWR	828.4257	CRHISR
837.4465	YEELRK	905.4662	SVMNWIR
876.4210	NFPDLDR	925.5003	IWGEHRK
900.4785	AASQPGEIK	927.5054	RCRHISR
904.4734	KAVDENTK	952.4894	NFTRRCR
908.4836	IIYNTNER	984.5268	RCRHISR
950.3850	HDDDSSFK	1009.5108	NFTRRCR
963.4782	FPNEDTLK	1091.5633	WPPPDQVPR
972.4165	MMSETLNF	1102.5813	DPMATLIVDK
991.5683	LLRQGNYK	1115.6307	SNLELEVALK
996.4534	TYWEQNR	1133.6235	SMLELAAATVK
1034.5742	VAVPHSPATR	1167.5501	NNAHDVDLNR
1073.5334	SNAQGVDLNR	1227.5528	YAPSPDDSTFK
1078.4800	KHDDDSSFK	1230.6762	KDPMATLIVDK
1136.5081	EELMDWWK	1235.6314	SVMNWIRSSR
1162.6691	KVAVPHSPATR	1261.7184	KSMLELAAATVK
1193.6062	DGDYWRLLR	1406.7890	YKSNLELEVALK
1271.6266	IYTIGESFEGR	1444.6638	YVGNMHGDEPLGR

It can be observed that sequence similarity does not equate to a match of mass spectra. The results obtained from the above described experiments motivated us to think more critically. This led us to design the bioinformatics pipeline that uses the limited information available, i.e. non-plant peptides reported by Pro Group, and to use them for protein identification.

All the algorithms to compare sequences rely on some scheme (evolutionary or physicochemical-based) to assign a score to each match. With the BLAST algorithm each AA position is independently scored during matching of protein sequences to get the best score/alignment. However, when considering a spectrum produced by MS/MS, individual AA position cannot be scored independently (isolated from one another). The spectrum reflects a combination of AA sequences. For example, changing the order of AAs in a peptide, or changing the AA which follows a spe-

cific residue may well significantly change the spectrum produced. Similarity-based searches, as commonly carried out, do not equate to similarity of the mass spectra.

In a BLAST search, the scoring matrix is an input parameter. Thus an evolution-oriented matrix can be replaced by one based on similarity of physico-chemical properties. Many such scoring matrices have been proposed [19, 22, 24, 26, 29, 46, 49]. A possible improvement on the preceding alignment-based methodology is to use a physicochemical-based scoring matrix along with a plant database. However, in addition to the limitation described in the preceding paragraph, there is a second limitation associated with this approach.

In order to facilitate the interpretation of the reported "hits" (search results), BLAST reports Expectation values or E-value and a score. Statistically, the E-value shows the number of alignments with at least that score that one expects to observe by chance [44]. The smaller the E-value, the more likely it is that the reported hit is significant (biologically meaningful). Since the proposed method involves the use of a different scoring matrix (other than evolutionary based), it remains undetermined whether the statistical basis for the E-value will still hold true. Therefore, even if a match is found, we will be unable to determine if it is statistically significant.

Because of the aforementioned reasons the prospect of using an alignment tool along with physicochemical-based scoring matrix was not further investigated and encouraged us to develop a novel method to recover plant proteins by using the non-plant peptides reported by the software.

5.3 Idea Explored but not Associated with the Working of the Pipeline

Various ideas and techniques were implemented during the development of the pipeline. Some of the ideas were unsuccessful, presenting problems too complex to be solved in the limited available time. Such information could be beneficial to the readers who are trying to design alternate or complementary techniques involving processed data from (protein) identification software based on MS or for those who would like to extend the research presented here.

After completion of the pipeline to the point of generating the initial plant list (Figure 3.11), various other ideas were investigated, implemented and tested with the goal of increasing the confidence level in the potentially identified plant proteins. The ideas were of two types: either to find supporting evidence to confirm the result or to use alternate techniques to identify plant proteins. These alternate techniques were based on information other than m/z values. One such technique is presented next.

5.3.1 Use of Amino Acid Compositions

Amino acid composition or coverage was one choice considered. The concept was to use AA coverage as a second criterion to give higher preference to those protein sequences which had AA composition more similar to the input peptide sequences (the non-plant proteins reported by Pro Group). This can be achieved by assigning a coverage score to each reported protein. The coverage score would represent the degree of sequence deviation between the reported plant protein and the input non-plant peptides, with a larger score indicating increased coverage deviation.

The coverage score was calculated for each peptide in the input data as the sum of the squared differences in percent AA composition between that peptide and the peptide in the database. In other words, the idea was to characterize the composition of the input non-plant peptide and observe how much of the composition is conserved in the proteins in the database. For example, consider a peptide GHAK. If this peptide is matched to a peptide ARAGLK from the database, coverage score would be approximately 1388.61. The calculation of the score is shown below:
Peptide sequence (P1) = GHAK, Total Length = 4
Percentage of each AA in P1: G = (1/4)*100, H = (1/4)*100, A = (1/4)*100, K = (1/4)*100
Peptide sequence (P2) = ARAGLK, Total Length = 6
Percentage of each AA in P2: A = (2/6)*100, R = (1/6)*100, G = (1/6)*100, L = (1/6)*100, K = (1/6)*100
Coverage Score = $(G_{P1}\text{-}G_{P2})^2+(H_{P1}\text{-}0_{P2})^2+(A_{P1}\text{-}A_{P2})^2+(K_{P1}\text{-}K_{P2})^2+(0_{P1}\text{-}R_{P2})^2+(0_{P1}\text{-}L_{P2})^2$
Similarly if GHAK is matched with ISIPIR, the coverage score would be 4999.33. In the first case, at least two AA were conserved, while in the second case, none of the AA were conserved as is evident by the increased coverage scores. The sequence GHAK shares more AAs with ARAGLK than with ISIPIR. The concept was adopted from other work, the software package AACompIdent from the expert protein analysis system, ExPASy[4] [51]. The coverage score will be calculated for each peptide associated with the proteins present in the evidence file and then will be used to give higher preference to those protein sequences which had AA composition more similar to the input peptide sequences.

The coverage scoring scheme was implemented and tested. Given an input list of peptides, a list of protein IDs and associated coverage scores were output. To determine the usefulness of the technique, the output list was compared with the corresponding initial plant list looking for proteins which occurred in both lists. After testing (data not shown) with different sets of peptides (known plant, known non-plant, unknown peptides) and with different threshold values (10, 50, 100, PPM) it was not clear that the AA composition based method (sequence coverage) can be used to report proteins from such a list. Different sets of proteins were reported each time during testing by the

[4]http://www.expasy.org/tools/aacomp/aacomp-doc.html

implemented method (no common proteins were found in the lists). Amino acid composition based method is currently being used with *de novo* based approaches [15] for protein identification. More investigation is needed to find ways (if any) in which AA composition based method could be used in parallel with MS/MS based approaches for protein identification.

There were a number of other ideas devised for improving the pipeline. The ideas involve properties of peptides, such as MW, charge state(s) of the peptides, count of AAs present in the peptides, and length of the peptides. Exploration of these ideas are left as future work as described in Chapter 6.

Chapter 6

Conclusions And Future Studies

The research presented in this thesis shows that m/z values, consideration of unique peptides and accounting for proteins with shorter sequences can be used to identify proteins when limited information is available, as it is the case when a list of non-plant proteins is reported as being present in a plant-derived sample. The information available was limited because the original input data was already processed by the Pro Group software. It was observed that with equal numbers of peptides per protein sensitivity and PPV of the pipeline increased (Section3.5). During testing both 10 and 100 PPM showed better results than 50 PPM (Section3.5). However, the most stringent threshold value (10 PPM) showed better results compared to 50 or 100 PPM (Chapter 4) with real data set. The remainder of this chapter discusses different ideas to improve the pipeline; how to make it more robust, provide statistical information, correct any bias and make use of other information (such as MWs) that is available to report plant proteins.

As described in Chapter 2 and Chapter 3, PTMs and miss-cleavage sites could also be responsible for the reporting of non-plant proteins. Software packages (available commercially or as open source) provide various facilities to the user to compensate for the shift in peaks in the mass spectrum caused by such processes. The bioinformatics pipeline lacks these features. Including in the current pipeline modules that take PTMs and miss-cleavage sites into account could help in determining plant peptides, given non-plant peptide sequences and predicted m/z values.

A module to compensate for PTMs would scan the peptide sequences present as input data as well as the peptide sequences present in the plant-only database. The pipeline would then try to compensate for the shift in the peaks (m/z values) caused by PTMs. Another module to compensate for miss-cleavage sites can also be added. In a sequence there can be more than one miss-cleavage site therefore setting a maximum number of miss-cleavage site(s) to allow might be appropriate. Increasing this maximum would increase the search space, making identification more time-consuming.

In Chapter 3 it was demonstrated that different rules are followed by different software programs to produce tryptic digests. In some cases, for the same protein many different peptides can be reported. Hence, if different tryptic digest software was used in the pipeline, potential m/z values would also change according to the different AA composition of the peptide sequences. The net

affect of this would be the reporting of different sets of plant proteins by the pipeline. A module can be added to the current pipeline that uses a different set of rules to produce tryptic digests. That is, another program can be used or a different module can be added to the pipeline to digest proteins into (tryptic) peptides instead of using digest from the EMBOSS software package to get tryptic digests as expected and use them to report plant proteins.

The current implementation of the pipeline uses m/z values of the peptides for protein identification. The MW associated with each input peptide can be also used in parallel to identify proteins or improve the confidence in the reported/identified proteins by the pipeline. The MWs of input peptides can be matched with the MWs of digested plant protein peptides controlled by a threshold value. Damodaran et al.[13] have proposed a "value-based scoring system". In this approach among other values MWs and isoelectric point of peptide sequences were used for protein identification. This paper would be a good starting point for further studies.

In Chapter 4 it was shown that the third potential plant list attempts to compensate for the bias in the pipeline towards heavier proteins observed in the initial plant list. From the results presented in Chapter 4 it was noticeable that in the third potential plant list, proteins with small sequences were reported among the top hits. In all such cases, proteins with sequence length in hundreds (100-500) or smaller were observed. Proteins with sequence length anywhere between 10-200 were most common. The pipeline has overly penalized longer, heavier proteins. One method to correct the bias is to calculate the total number of theoretical tryptic peptides and divide this number by the length of the protein in order to normalize the lengths of the peptides.

The pipeline lacks any statistical estimation of accuracy or error rate in protein identification. A probability-based approach can be investigated and implemented to report such values for proteins reported by the pipeline.

References

[1] A. Bleasly, Personal Communication, EMBOSS support, March 3rd, 2008.

[2] R. Aebersold and D.R. Goodlett. Mass Spectrometry in proteomics. *American Chemical Society*, 101:269–295, 2001.

[3] R. Aebersold and M. Mann. Mass Spectrometry-based proteomics. *Nature*, 422:198–207, 2003.

[4] K. Aggarwal, L.H. Choe, and K.H. Lee. Shotgun proteomics using the iTRAQ isobaric tags. *Briefings in Functional Genomics and Proteomics*, 5(2):112, 2006.

[5] P. Baker and K. Clauser. University of California San Francisco Mass Spectrometry Facility, 2006. ProteinProspector, URL: http://prospector.ucsf.edu/, Last Visited November 16, 2006.

[6] R.C. Beavis and D. Fenyö. Database searching with mass-spectrometric information. *Trends in Biotechnology*, 18:22–27, 2000.

[7] P. Berndt, U. Hobohm, and H. Langen. Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. *Electrophoresis*, 20:3521–3526, 1999.

[8] Applied Biosystems. *Using Pro Group Reports*. Applied Biosystems/MDS Sciex, 2006.

[9] Applied Biosystems. ProteinPilot software for protein identification and expression analysis. Technical Note, Applied Biosystems/MDS Sciex, 2006. URL: http://www3.appliedbiosystems.com/cms/groups/psm_marketing/documents/generaldocuments/cms_042621.pdf, Last accessed January 2007.

[10] K. Boutilier, M. Ross, A.V. Podtelejnikov, C. Orsi, R. Taylor, P. Taylor, and D. Figeys. Comparison of different search engines using validated MS/MS test datasets. *Analytica Chimica Acta*, 534(1):11–20, 2005.

[11] D.C. Chamarad, G. Körting, K. Stühler, H.E. Meyer, J. Klose, and M. Blüggel. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*, 4(3):619–628, 2004.

[12] P.K. Chong, C.S. Gan, T.K. Pham, and P.C. Wright. Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: Implication of multiple injections. *Journal of Proteome Research*, 5(5):1232–40, 2006.

[13] S. Damodaran, T.D. Wood, P. Nagarajan, and R.A. Rabin. Evaluating Peptide Mass Fingerprinting-based Protein Identification. *Genomics, Proteomics & Bioinformatics*, 5(3-4):152–157, 2007.

[14] J.B. Fenn, M. Mann, C.K. Meng, S.F. Wong, and C.M. Whitehouse. Electrospray ionization-principles and practice. *Mass Spectrometry Reviews*, 9(1):37, 1990.

[15] F. Forner, L.J. Foster, and S. Toppo. Mass spectrometry data analysis in the proteomics era. *Current Bioinformatics*, 2(1):63–93, 2007.

[16] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, and A. Bairoch. Protein identification and analysis tools on the ExPASy server. *The proteomics protocols handbook*, pages 571–607, 2005.

[17] M.C. Giddings, M.R. Holmes, and Ramkissoon K. Proteomics and Protein Identification. In Baxevanis A.D. and Ouellette B.F.F., editors, *Bioinformatics: A practical guide to the analysis of genes and proteins*, chapter 18. John Wiley and Sons, Inc, 3rd edition, 2005.

[18] S.P. Gygi, S.A. Gerber, F. Turecek, M.H Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, 17:994–999, 1999.

[19] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–10919, 1992.

[20] P. Jeno. Protein identification. Lecture Notes. www.biozentrum.unibas.ch/jeno/index.html. Last accesses on December 2007.

[21] J.V. Jorrín, A.M. Maldonado, and M.A. Castillejo. Plant proteome analysis: a 2006 update. *Proteomics*, 7(16):2947–2962, 2007.

[22] A. Kidera, Y. Konishi, M. Oka, T. Ooi, and H.A. Scheraga. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *The Protein Journal*, 4(1):23–55, 1985.

[23] F. Kolderup. Effects of temperature, photoperiod, and light quantity on protein production in wheat grains. *Journal of the Science of Food and Agriculture*, 26(5):583–592, 2006.

[24] J.M. Koshi and R.A. Goldstein. Mutation matrices and physical-chemical properties: Correlations and implications. *Proteins Structure Function and Genetics*, 27(3):336–344, 1997.

[25] D.C. Liebler. *Introduction to Proteomics: Tools for the New Biology*. Humana Press, New Jersey, 2002.

[26] X. Liu and W. Zheng. An amino acid substitution matrix for protein conformation identification. *Journal Of Bioinformatics And Computational Biology*, 4(3):769–782, 2006.

[27] M. Mann and O.N. Jensen. Proteomic analysis of post-translational modifications. *Nature Biotechnology*, 21(3):255–261, 2003.

[28] R. Matthiesen. Methods, algorithms and tools in computational proteomics: a practical point of view. *Proteomics*, 7(16):2815–2832, 2007.

[29] K. Nakai, A. Kidera, and M. Kanehisa. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Engineering Design and Selection*, 2(2):93–100, 2002.

[30] A.I. Nesvizhskii and R. Aebersold. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discovery Today*, 9(4):173–181, 2004.

[31] O. Nørregaard Jensen. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Current Opinion in Chemical Biology*, 8(1):33–41, 2004.

[32] A. Pandey and M. Mann. Proteomics to study genes and genomes. *Nature*, 405(6788):837–846, 2000.

[33] D.N. Perkins, D.J. Pappin, D.M. Creasy, and J.S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[34] G.E. Reid and S.A. McLuckey. "Top down"protein characterization via tandem mass spectrometry. *Journal of Mass Spectrometry*, 37(7):663–675, 2002.

[35] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European molecular biology open software suite. *Trends in Genetics*, 16(6):276–277, 2000.

[36] J. Rodriguez, N. Gupta, R.D. Smith, and P.A. Pevzner. Does trypsin cut before proline? *Journal of Proteome Research*, 7(1):300–305, 2008.

[37] P.L. Ross, Y.N. Huang, J.N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, et al. Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging Reagents. *Molecular and Cellular Proteomics*, 3(12):1154–1169, 2004.

[38] M. Schneider, M. Tognolli, and A. Bairoch. The Swiss-Prot protein knowledgebase and ExPASy: providing the plant community with high quality proteomic data and tools. *Plant Physiology and Biochemistry*, 42(12):1013–1021, 2004.

[39] I.P. Shadforth, T.P. Dunkley, K.S. Lilley, and C. Bessant. i-Tracker: for quantitative proteomics using iTRAQ. *BMC Genomics*, 6(1):145, 2005.

[40] J.C. Silva, M.V. Gorenstein, G.Z. Li, J.P.C. Vissers, and S.J. Geromanos. Absolute quantification of proteins by LCMSE a virtue of parallel MS acquisition. *Molecular and Cellular Proteomics*, 5(1):144–156, 2006.

[41] D.J. Skylas, D. Van Dyk, and C.W. Wrigley. Proteomics of wheat grain. *Journal of Cereal Science*, 41(2):165–179, 2005.

[42] H. Steen and M. Mann. The abc's (and xyz's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5(9):699–711, 2004.

[43] W.A. Tao and R. Aebersold. Advances in quantitative proteomics via stable isotope tagging and mass spectrometry. *Current Opinion in Biotechnology*, 14(1):110–118, 2003.

[44] A.T. Tatiana and L.M. Thoams. BLAST 2 sequence, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters*, 174:247–250, 1999.

[45] K.B. Tomer, F.W. Crow, and M.L. Gross. Location of double-bond position in unsaturated fatty acids by negative ion MS/MS. *Journal of the American Chemical Society*, 105(16):5487–5488, 1983.

[46] K. Tomii and M. Kanehisa. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Engineering Design and Selection*, 9(1):27–36, 2001.

[47] R.M. Twyman. *Principles of Proteomics*. BIOS Scientific Publishers, New York, 2004.

[48] M. Tyers and M. Mann. From genomics to proteomics. *Nature*, 422(6928):193–197, 2003.

[49] M.S. Venkatarajan and W. Braun. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. *Journal of Molecular Modeling*, 7(12):445–453, 2001.

[50] M. Vihinen. *Bioinformatics in Proteomics*. New Jersey, 2003.

[51] M.R. Wilkins, E. Gasteiger, A. Bairoch, J.C. Sanchez, K.L. Williams, R.D. Appel, and D.F. Hochstrasser. Protein identification and analysis tools in the ExPASy Server. In *2-D Proteome Analysis Protocols*. Humana Press, New Jersey, 1998.

[52] J.R. Yates. Mass spectrometry and the age of the proteome. *Journal of Mass Spectrometry*, 33(1):1–19, 1998.

[53] J.R. Yates. Mass spectrometry from genomics to proteomics. *Trends in Genetics*, 16(1):5–8, 2000.

[54] L. Zamdborg, R.D. LeDuc, K.J. Glowacz, Y.B. Kim, V. Viswanathan, I.T. Spaulding, B.P. Early, E.J. Bluhm, S. Babai, and N.L. Kelleher. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Research*, 35(Web Server issue):W701, 2007.

Appendix A

N	Unused ProtSc	Total ProtSc	Accession	Protein Name
131	1.30	1.30	T03584	plastocyanin precursor [validated] - rice
132	1.30	1.30	S53749	histone H4 - rat
133	1.30	1.30	Q87RP7_VIBPA	Putative hemolysin.- Vibrio parahaemolyticus.
134	1.30	1.30	Q6YZK5_ORYSA	Hypothetical protein P0488B06.35.- Oryza sativa (japonica cultivar-group).
135	1.30	1.30	Q9EQB4_MOUSE	Odorant receptor K15 (Olfactory receptor Olfr937) (Olfactory receptor MOR171-24).- Mus musculus (Mou
136	1.30	1.30	AAS04869	AE016958 NID: - Mycobacterium avium subsp. paratuberculosis str. k10
137	1.30	1.30	Q5SCZ1_HUPLU	NADH-plastoquinone oxidoreductase subunit 8.- Huperzia lucidula (Shining clubmoss) (Lycopodium lucid
138	1.30	1.30	Q5T4F0_HUMAN	OTTHUMP00000040262.- Homo sapiens (Human).
139	1.30	1.30	Q7Z659_HUMAN	Hypothetical protein DKFZp779H0156 (Fragment).- Homo sapiens (Human).
140	1.30	1.30	Q61P36_CAEBR	Hypothetical protein CBG07723.- Caenorhabditis briggsae.
141	1.30	1.30	Q6Z340_ORYSA	Hypothetical protein B1121A12.19.- Oryza sativa (japonica cultivar-group).
142	1.30	1.30	Q6AVQ3_ORYSA	Hypothetical protein OSJNBa0075M12.19.- Oryza sativa (japonica cultivar-group).
143	1.30	1.30	Q9J2H5_9GAMA	ORFRU12-L.- Rhesus monkey rhadinovirus H26-95.
144	1.30	1.30	Q6P4K3_XENTR	Hypothetical protein MGC75963.- Xenopus tropicalis (Western clawed frog) (Silurana tropicalis).
145	1.30	1.30	Q6YV17_ORYSA	Hypothetical protein OSJNBb0060O16.31.- Oryza sativa (japonica cultivar-group).
146	1.30	1.30	Q7XBY4_ORYSA	Putative carnitine/acylcarnitine translocase.- Oryza sativa (japonica cultivar-group).
147	1.30	1.30	Q89MW2_BRAJA	Transcriptional regulatory protein.- Bradyrhizobium japonicum.

Figure A.1: Figure displaying a Pro Group report. The report order is shown under the heading N. This is followed by the two scores. The column headed Unused ProtSc shows the *Unused score* and Total ProSc shows the *Total score*. The accession numbers of the associated proteins are shown under the column heading Accession followed by the protein descriptions under heading Protein Name. Figure adapted from [8].

Table A.1: Calculations of sensitivity and positive predictive value for each test case shown in Table 3.5. Better performing list or lists are underlined. In the table entry "NC" indicates a value that was non-calculable. For example, in test case 4 TPs and FNs have zero values. Therefore sensitivity cannot be calculated.

Sensitivity	**Positive Predictive Value**
Test 1 with 100 PPM	
List 1: TP = 2, FN = 1, 2/(2+1) (66%)	List 1: TP = 2, FP = 28, 2/(2+28) (6.6%)
List 2: TP = 2, FN = 1, 2/(2+1) (66%)	List 2: TP = 2, FP = 28, (2/2+28) (6.6%)
List 3: TP = 1, FN = 2, 1/(1+2) (33%)	List 3: TP= 1, FP = 29, 1/(1+29) (3.3%)
Test 1 with 50 PPM	
List 1: TP = 0, FN = 3, 0%	List 1: TP = 0, FP = 30, 0%
List 2: TP = 1, FN = 2, 33%	List 1: TP = 1, FP = 29, 3%
List 3: TP = 1, FN = 2, 33%	List 1: TP = 1, FP = 29, 3%
Test 1 with 10 PPM	
List 1: TP = 2, FN = 1, 66%	List 1: TP = 2, FP = 28, 6.6%
List 2: TP = 2, FN = 2, 50%	List 1: TP = 2, FP = 28, 6.6%
List 3: TP = 1, FN = 2, 50%	List 1: TP = 1, FP = 29, 3%
Test 2 with 100 PPM	
List 1: TP = 4, FN = 0, 100%	List 1: TP = 4, FP = 26, 13%
List 2: TP = 4, FN = 0, 100%	List 2: TP = 2, FP = 26, 7%
List 3: TP = 1, FN = 3, 25%	List 3: TP = 1, FP = 29, 3%
Test 2 with 50 PPM	
List 1: TP = 1, FN = 3, 25%	List 1: TP = 1, FP = 29, 3%
List 2: TP = 2, FN = 2, 50%	List 2: TP = 2, FP = 28, 6.6%
List 3: TP = 1, FN = 3, 25%	List 3: TP = 1, FP = 29, 3%
Test 2 with 10 PPM	
List 1: TP = 4, FN = 0, 100%	List 1: TP = 4, FP = 26, 13%
List 2: TP = 4, FN = 0, 100%	List 2: TP = 4, FP = 26, 13%
List 3: TP = 2, FN = 2, 50%	List 3: TP = 2, FP = 28, 6.6%
Test 3 with 100 PPM	
List 1: TP = 3, FN = 1, 75%	List 1: TP = 3, FP = 27, 10%
List 2: TP = 3, FN = 1, 75%	List 2: TP = 3, FP = 27, 10%
List 3: TP = 0, FN = 4, 0%	List 3: TP = 0, FP = 30, 0%
Test 3 with 50 PPM	
List 1: TP = 0, FN = 4, 0%	List 1: TP = 0, FP = 30, 0%
List 2: TP = 1, FN = 3, 25%	List 2: TP = 1, FP = 29, 3.3%
List 3: TP = 0, FN = 4, 0%	List 3: TP = 0, FP = 30, 0%
Test 3 with 10 PPM	
List 1: TP = 3, FN = 0, 100%	List 1: TP = 3, FP = 27, 10%
List 2: TP = 3, FN = 0, 100%	List 2: TP = 3, FP = 27, 10%
List 3: TP = 2, FN = 1, 6.6%	List 3: TP = 2, FP = 28, 6.6%
Test 4 with 100 PPM	
NC	List 1: FP = 30, 0%
NC	List 2: FP = 30, 0%
NC	List 3: FP = 30, 0%
Test 5 with 100 PPM	
NC	List 1: FP = 30, 0%
NC	List 2: FP = 30, 0%
NC	List 3: FP = 30, 0%

Table A.2: Calculations of sensitivity and positive predictive value for each test case shown in Table 3.7. In the table entry "NC" stands for a non-computable value. For example, in test case 4 TPs and FNs have zero values. Therefore sensitivity cannot be calculated.

Sensitivity	**Positive Predictive Value**
Test 1 with 100 PPM	
TP = 2, FN = 1, 2/(2+1) (66%)	TP = 2, FP = 11, 2/(2+11) (15%)
Test 1 with 50 PPM	
TP = 1, FN = 2, 33%	TP = 1, FP = 16, 5.8%
Test 1 with 10 PPM	
TP = 2, FN = 1, 66%	TP = 2, FP = 21, 8.6%
In test 1, 100 and 10 PPM showed better results	100 PPM was better than 50 or 10 PPM.
Test 2 with 100 PPM	
TP = 4, FN = 0, 100%	TP = 4, FP = 8, 50%
Test 2 with 50 PPM	
TP = 2, FN = 2, 50%	TP = 2, FP = 10, 16.6%
Test 2 with 10 PPM	
TP = 4, FN = 0, 100%	TP = 4, FP = 23, 14.8%
In test 1, 100 and 10 PPM showed better results	100 PPM was better than 50 or 10 PPM.
Test 3 with 100 PPM	
TP = 3, FN = 0, 100%	TP = 3, FP = 15, 16.6%
Test 3 with 50 PPM	
TP = 0, FN = 3, 0%	TP = 0, FP = 23, 0%
Test 3 with 10 PPM	
TP = 3, FN = 0, 100%	TP = 3, FP = 21, 12.5%
In test 3, 100 and 10 PPM showed better results	100 PPM was better than 50 or 10 PPM.
Test 4 with 100 PPM	
TP = 0, FN = 0, NC	TP = 0, FP = 6, 0%
Test 5 with 100 PPM	
TP = 0, FN = 0, NC	TP = 0, FP = 15, 0%

Unused ProtSc	Total ProtSc	Accession	Protein Name	Species	Conf	Sequence
15.00	15.00	PWWTB	H+-transporting two-sector ATPase (EC 3.6.3.14) beta chain - wheat chloroplast		1	AFSPGJ
0.00	13.70	Q6ENG7_ORYNI	ATPase beta subunit.- Oryza nivara (Indian wild rice).		99	AHGGLSVFGGVGER
0.00	13.00	BAA01870	AEGATPS1 NID: - Aegilops columnaris		0	AHGGVSIFGGVGER
0.00	13.00	AAA84588	RICCPCTA NID: - Oryza sativa		2	AHR
0.00	13.00	Q9MVP6_ORYSA	ATP synthase beta subunit.- Oryza sativa (Rice).		99	AITLEEENJ
0.00	13.00	PWBHB	H+-transporting two-sector ATPase (EC 3.6.3.14) beta chain - barley chloroplast		1	AJAAJ
0.82	11.82	Q9TMR4_9ASPA	ATP synthase beta subunit (Fragment).- Hyacinthus orientalis.		1	DTVGJ
0.00	11.80	Q9TLL9_USEUD	ATP synthase beta subunit (Fragment).- Corydalis nobilis.		1	EKAIJ
0.00	11.70	Q6ENV6_SACOF	ATP synthase beta subunit.- Saccharum officinarum (Sugarcane).		98	ESGVINEJ
0.00	11.00	AAP53257	AE016959 NID: - Oryza sativa (japonica cultivar-group)		5	FPPGJ
0.00	11.00	Q9TMP1_9LILI	ATP synthase beta subunit (Fragment).- Tapeinochilos ananassae.		99	FVQAGSEVSALLGR
0.00	11.00	Q9TMS7_9ASPA	ATP synthase beta subunit (Fragment).- Eucharis grandiflora.		99	GIYPAVDPLDSTSTMLQLR
0.00	11.00	Q9TMR6_9LILI	ATP synthase beta subunit (Fragment).- Helmholtzia acorifolia.		15	GMEVIDTGSPLSVPVGGATLGR
0.00	11.00	Q9TMR5_HEMLI	ATP synthase beta subunit (Fragment).- Hemerocallis lilioasphodelus (Yellow daylily).		15	GMEVVDTGTPLSVPVGGATLGR
0.00	11.00	Q9TMQ3_9LILI	ATP synthase beta subunit (Fragment).- Pandanus utilis.		99	IDQIIGPVLDVTFPPGJ
0.00	11.00	Q9TMP3_9LILI	ATP synthase beta subunit (Fragment).- Stichoneuron caudatum.		99	IGLFGGAGVGJ
0.00	11.00	Q9TMS4_9POAL	ATP synthase beta subunit (Fragment).- Glomeropitcairnia penduliflora.		68	IVGEKHYETAQR
0.00	11.00	Q9TMM7_9ROSI	H(+)-transporting ATP synthase (EC 3.6.1.34) (Fragment).- Muntingia calabura.		68	IVGKEHYETAQR
0.00	11.00	Q9TMN8_9ASPA	ATP synthase beta subunit (Fragment).- Xanthorrhoea quadrangulata.		68	IVGQEHYETAQR
0.00	11.00	Q9TMR3_9ASPA	ATP synthase beta subunit (Fragment).- Liriope muscari (big blue lilyturf).		18	IVSNEHYETAQR
0.00	11.00	Q9TMR2_MARBI	ATP synthase beta subunit (Fragment).- Maranta bicolor.		52	KDVLLFIDNIFR
0.00	11.00	Q9XPT2_9ROSI	H(+)-transporting ATP synthase (EC 3.6.1.34) (Fragment).- Aquilaria beccariana.		99	LSIFETGIJ
0.00	11.00	Q9TMP6_9ASPA	ATP synthase beta subunit (Fragment).- Ledebouria socialis.		1	PGVSALENKNLGR
0.00	11.00	Q9TMQ8_MUSAC	ATP synthase beta subunit (Fragment).- Musa acuminata (Banana).		1	PPGJ
0.00	11.00	Q9TMQ9_9LILI	ATP synthase beta subunit (Fragment).- Monocostus uniflorus.		99	SAPAFIELDTJ
0.00	11.00	Q9TMQ1_PHEGU	ATP synthase beta subunit (Fragment).- Phenakospermum guyannense (South American travelers palm).		2	SAPAFIQLETJ
0.00	11.00	Q9MT58_9LILI	ATP synthase beta subunit (Fragment).- Androcymbium ciliolatum.		85	TNPTTSGSTVSTLEEJ
0.00	11.00	Q9TMR1_9LILI	ATP synthase beta subunit (Fragment).- Marantochloa atropurpurea.		99	TNPTTSPPGASTIEEJ
0.00	11.00	Q9TMR7_HELRS	ATP synthase beta subunit (Fragment).- Heliconia rostrata (Lobster claw).		95	VAIVYGQMNEPPGAR
0.00	11.00	Q9TMT4_9LILI	ATP synthase beta subunit (Fragment).- Cyclanthus bipartitus.		95	VALVYGQMNEPPGAR
0.00	11.00	Q9TMT3_9LILI	ATP synthase beta subunit (Fragment).- Cymbocarpa refracta.		2	VISLJ
0.00	11.00	Q9TMT1_9LILI	ATP synthase beta subunit (Fragment).- Dasypogon bromeliifolius.		68	VVDLLAPYKR
0.00	11.00	Q9TMS9_9LILI	ATP synthase beta subunit (Fragment).- Dimerocostus strobilaceus.		99	VVDLLAPYR

Figure A.2: Figure displaying peptides and other information in the Pro Group report. From left-to-right the figure displays the two scores used by Pro Group for the proteins identified by Pro ID. The column headed Unused Pror Sc and TotalProtSc contain the two scores. The accession IDs associated with the proteins are shown under the heading Accession. Proteins names are shown under Protein Name. The column heading Species and Conf are used internally by Pro Group. The peptide sequences are shown under the column heading Sequences. Selecting the accession number will highlight the associated peptides. This is a feature of a Pro Group report. Figure adapted from [8].

EPVSGSLLYGNNIISGAIIPTSAAIGLHFYPIWEAASVDEWLYNGGPYELIVLHFLLGVACY
MGR
ESTSLWGRFCNWITSTENRLYIGWFGVLMIPTLLTATSVFIIAFIAAPPVDIDGIR
SLHFFLAAWPVVGIWFTALGISTMAFNLNGFNFNQSVVDSQGR
ANLGMEVMHER
WYLINFWQYFFSFWTQPRRIHLNQLANSCFDFLGYLSSVPK
SIHSIFPFLEDKFLHLDYLSHIEIPYPIHFEILVQLLQYRIK
MYQQNFWINSVNHPNQDQLLDYKIGFYSEFYSQILPEGFAIVVEIPFSLR
SIHSIFPFLEDKFLHLDYLSHIEIPYPIHFEILVQLLQYRIK
QSSSLPLLSSGTFLERIIFSRK
TIWFFMDPLMHYVR
DVPSLHLLRFFLNYYSNWNSFITSMK
SIFLFKK
TSLFSFR
MEHFGIMYPGFFRK
TLTHTSNLYHTEPGATLARKLTATSFADR
VAAVMAAGDHGSTFAGGPLVCAVANEVFDR
TCAVFVEPVQGEGGIYPADAEFLR
VFFCNSGTEANEGALKFARK
APFAPGLPGNTFTPYGDLEK
NAGDAAWETVSFENGFHGR
GDILRLVPPLIVSSAQVQK
SSLTEKLAGHPRLR
VSDPAFLANVTER
TFGALSLTWK

Figure A.3: Non-plant peptides used as input for test 1. A total of 24 peptides were used as input.